*Journal of the*
# OPTICAL SOCIETY *of* AMERICA

VOLUME 35, NUMBER 3 MARCH, 1945

## The Spectral Sensitivity of the Human Eye

### I. A Spectral Adaptometer

GEORGE WALD*
*Biological Laboratories, Harvard University, Cambridge*
(Received December 28, 1944)

TO explore the spectral sensitivity of human vision under a wide variety of conditions, a more adaptable and portable apparatus was wanted than is ordinarily employed in such measurements. The description of such an instrument follows. It provides a number of highly monochromatic lights in broad ranges of intensity and spaced widely throughout the spectrum; and also arrangements for regulating the size of the test field, its exposure, and the location of its image on the retina. Several general matters involved in the design and use of this instrument also are discussed—properties of the high-pressure mercury arc, the isolation of mercury radiations with light filters, and the use of the Beckman spectrophotometer for low intensity photometry. A special problem—the test of color filters for radiation leakage—is treated separately in an Appendix.

#### DESIGN

The construction of the instrument is shown in Fig. 1. Its central characteristic is the use of a high pressure mercury arc as source. From the mercury radiation ten highly monochromatic regions of the spectrum, ranging from 365 to 750 mμ, are isolated with color filters. The intensity of each of these radiations is regulated accurately with a neutral wedge. A shutter is provided for fixing the duration of flash to which the observer's eye is exposed. A variable fixation point is also provided, by means of which the image of the test field may be located in any desired area of retina within a radius of 12° of the fovea. Every effort has been made to keep the instrument rugged and portable. No difficulty has been experienced in transporting it by auto for long distances, and setting it up for immediate use wherever a dark room has been available.

The lamp which has been used in this adaptometer is a capillary mercury arc, the quartz inner capsule of the General Electric H-3 lamp. The vertical rods which support the capsule have been cut away from the conventional base, and are held in brass lugs in a cast-aluminum housing.[1] This is an 85-watt lamp, operated as are all lamps of this type from a special transformer, run in turn for greater stability from a voltage stabi-

* This research was supported in part by the Josiah Macy, Jr., Foundation of New York City. Short descriptions of the apparatus have already been published elsewhere [G. Wald and H. M. Burian, Am. J. Ophthal. **27**, 950 (1944); G. Wald, J. Opt. Soc. Am. **34**, 769 (1944)].

[1] The arrangement of the lamp in its housing, fitted with a focusing tube and with a housed transformer, is manufactured by the Mico Instrument Company of Cambridge, Massachusetts.



BOSTON UNIVERSITY
COLLEGE OF LIBERAL ARTS
LIBRARY

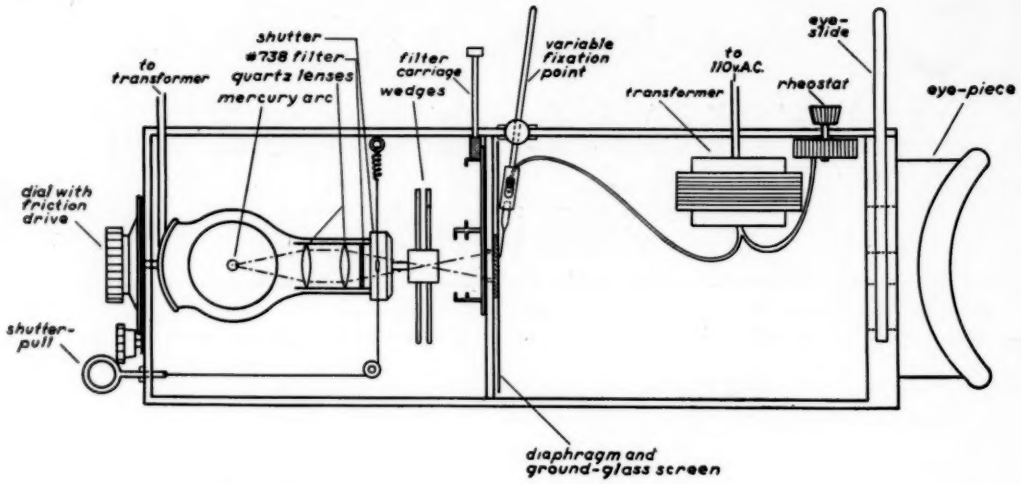



FIG. 1. Top view of the adaptometer, drawn approximately to scale.

lizer. The description which follows is concerned specifically with this arrangement.[2]

The barrel of the lamp housing holds a pair of quartz lenses which focus the arc discharge on the wedges; and in addition a Corning 738 filter to eliminate wave-lengths below about 350 mμ, which might otherwise permanently discolor the wedges.

To the end of the lamp housing is attached a photographic shutter. This can be set for "bulb" or time exposures, or to deliver flashes of graded durations from 1/100 to 1 second. The shutter is tripped by a string which runs to a pull-ring at the rear of the adaptometer. A spring attached to the adjacent wall helps to return the trip-lever to its position of rest.

The wedges in the instrument are of annular form and were made by the graded exposure of photographic plate.[3] A mated pair of wedges rotates in opposite directions, each wedge compensating the other. The total intensity range so provided is about six logarithmic units (about 1,000,000 times). At the rear of the instrument, on the same shaft with the wedges, is mounted a rotary dial from which the wedge position, and hence the intensity setting is read.

Just in front of the wedges a sliding carriage is provided which holds two color filters. This permits the rapid comparison of visual responses in two isolated regions of the spectrum. An extra channel allows a neutral or color filter to be added when desired to one of the regular filter combinations.

The light having passed through the wedge and neutral and color filters, impinges on a screen of Corex glass, ground on both faces. This is mounted behind a sheet-metal shield containing an opening for the test field. On the shield is drawn a grid of concentric circles and radiating lines, to guide the location of the fixation point at various distances from the center of the field. The arrangement is somewhat comparable to tangent screens used in visual perimetry. The maximum diameter of the field subtends a visual angle of 2°. Small diaphragms may be placed over the 2° opening to reduce its size to any desired dimension.

[2] Recently, because of its greater availability, I have mounted the inner capsule of a General Electric H-4 mercury lamp in the same housing. This is a 100-watt lamp, the radiation of which resembles that of the H-3. My experience with the H-4 capsule is still too short to judge the relative lifetimes of the two lamps, operated without the protection of their glass envelopes. The adaptometer might well be re-arranged to use the H-4 lamp in its normal format, since the glass envelope transmits the 365 mμ radiation very well, and the instrument is not otherwise designed to employ lower wave-lengths.

[3] The wedges were among a number manufactured about twenty years ago at the Jefferson Physical Laboratories of Harvard University with a specially designed device built by Mr. D. W. Mann. They have been described by Waite, Derby, and Kirk, Trans. Ophthal. Soc. U. Kingdom, **45**, 301 (1925).

The variable fixation point is constructed following a principle described by Abney and Watson (1915).[4] A midget lamp is mounted at the broad end of a piece of Pyrex rod, the other end of which is drawn out to a long, slender tip. Light is conducted by internal reflection from the lamp down the length of the rod to form a bright point at its tip end. The rod and lamp are mounted in a narrow brass tube at the end of a long brass stem. The latter is held in a ball joint so as to be freely movable in all directions. In this way the illuminated tip may be set at any position within a radius of 12° from the center of the field. Since the glass rod is transparent toward the tip, the latter may be placed at the center of the field without interfering in any way with its visibility. The midget lamp is operated from a 6-volt transformer, and is in series with a rheostat by means of which the brightness of the fixation point can be regulated at will.

The field and fixation point are enclosed in a viewing section which forms the front end of the adaptometer. This holds a rubber-cushioned eyepiece, wide enough to accommodate a subject's own spectacles when a lens correction is needed, and a sliding diaphragm which occludes either eye. All measurements are therefore monocular.

## MERCURY LAMP

As this description is written, a single lamp has been in use in the adaptometer for almost two years. It has yielded a remarkably uniform and constant radiation. It has also proved to be remarkably rugged mechanically, having been subjected repeatedly to the shocks and vibration of long distance transportation by auto.

In the commercial literature which describes this lamp, it is stated to require from 2 to 3 minutes to attain full brilliance. This is not an accurate description of its behavior. At most wave-lengths the intensity rises rapidly for about ten minutes. Later changes are comparatively small, but regular and reproducible. The course of these variations, measured with a photo-cell at a number of wave-lengths, is shown in Fig. 2. It is curious that the intensity of the ultraviolet radiation at 365 mμ rises to a maximum in the first 7 to 10 minutes, and then gradually falls. The intensity at other wave-lengths rises gradually for as long as 50 to 60 minutes. The extreme changes from 20 to 60 minutes after the lamp has been started occur at 365 mμ and 546 mμ. These are respectively a decrease in intensity of 0.05 log unit and a rise of 0.03 log unit. Such changes are negligible in visual adaptometry. In all work done with this instrument the lamp has been run for at least 30 minutes before beginning either calibrations or visual measurements.

## WEDGE CALIBRATION

The way in which the wedge decreases the intensity of the test field has been calibrated at each of the wave-lengths investigated. As noted above, the wedge used was made of photographic material. This of course is granular in composition, and its absorption of light is accompanied by considerable scattering. It must be calibrated, therefore, in the light path of the

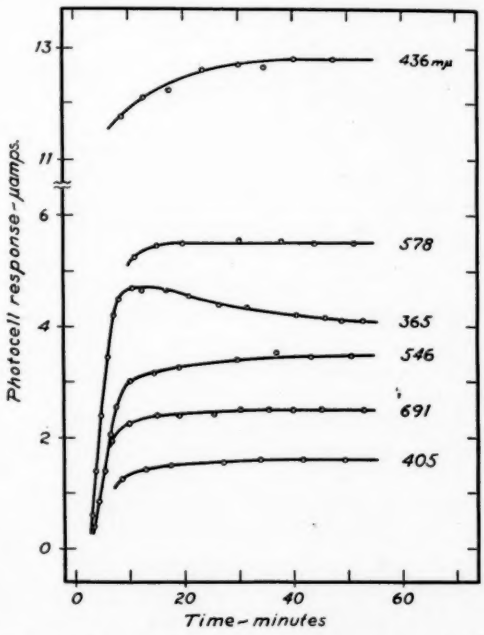



FIG. 2. Changes in the intensity of isolated portions of the mercury-arc radiation as a function of time after turning on the lamp (the inner capsule of the General Electric H-3 Lamp). The intensities were measured with a barrier-layer photo-cell placed in the position of the test field of the adaptometer. Ordinates show the response of the photo-cell in microamperes.

[4] W. de W. Abney and W. Watson, Phil. Trans. Roy. Soc. London, **A216**, 91 (1915).

adaptometer, with the calibrating device placed in the position of the test field.

The amounts of monochromatic radiation transmitted by such narrow band-pass filters as were used in this instrument are small at best; and since these are decreased by the wedge itself ultimately to $10^{-6}$, a very sensitive device must be used to obtain accurate intensity measurements. The photometer of a Beckman spectrophotometer[5] is admirably suited to this purpose. The photo-cell compartment, which contains photoelectric cells sensitive to the ultraviolet and visible regions, and to the red and infra-red, is easily disengaged from the body of the instrument, to which it remains attached only by a flexible cable. The instrument can therefore be used as a photoelectric photometer which, with its built-in amplifier, is sufficiently sensitive to measure such low intensities as are employed in adaptometry. The procedure is as follows:

The photo-cell compartment is detached from the instrument, rotated through 180°, and placed inside the adaptometer facing the test field. The mercury lamp is allowed to come to stable intensity, and appropriate color filters are inserted to isolate the wave-length region in which the wedge is to be calibrated. The Beckman spectrophotometer is designed to measure transmissions of some absorbing material relative to a "blank" of higher transmission. In the present instance the measurements are made by using a comparatively transparent portion of the wedge as blank, and determining the transmission of denser portions relative to it. Thus, relative to the transmission of the wedge when entirely open, denser portions are calibrated down to about 0.1 percent transmission. Then, relative to such a denser portion, the calibration is pursued further. In this way the total calibration of the wedge is pieced out.

This was done at each of the wave-lengths isolated in the instrument. Fortunately it turned out that over the entire portion of the wedge used in our measurements, the calibration of relative wedge densities was the same for all spectral regions. A single calibration curve could therefore be used for our measurements at all wave-lengths.

At a number of wave-lengths at which the mercury radiation is very bright, it is necessary to use in addition to the wedge a neutral filter of approximately 1 percent transmission. This was calibrated spectrophotometrically throughout the near ultraviolet and visible spectrum.

TABLE I. The isolation of mercury arc radiations with color filters. The filters are characterized in the first column by their wave-length "centers of luminosity." Their composition is shown in the second column; C. = Corning, J. = Jena (Schott und Gen.), W. = Wratten (Eastman). In the third column is shown the radiation of the high-pressure mercury arc transmitted by the filters, as viewed visually in a spectroscope, or—in the case of the 365 filter—a quartz spectrograph.

| Filter (mμ) | Composition | Spectroscopic transmission of Hg radiation |
|---|---|---|
| 365 | C. 5860, 738; J. *BG*12 1/2 | Lines at 365–366 mμ. |
| 405 | C. 3060, 4305, 5970 | Bright line at 405, dim line at 408 mμ |
| 436 | C. 3389, 5113 | Very bright line at 436, dim lines at 434, 435 mμ; dim continuum at 436–438 mμ. |
| 492 | W. 75, 48 | Brighter line at 492 mμ, dimmer at 496 mμ; dim continuum about 471–490 mμ. |
| 546 | W. 74, C. 5120 | Very intense line at 546 mμ, flanked by dim continuum to 549, very dim continuum to 536 mμ. |
| 578 | C. 4303, 3480 | Intense lines at 577, 579 mμ. |
| 621 | C. 2412, 9780 | Bright line at 623, dim line at 612 mμ. Continuum from about 609–635 mμ. |
| 691 | W. 91, C. 3962 | Bright line at 691 mμ, dim continuum from about 670–710 mμ. |
| 713 | W. 88, C. 3962 | Line at 707 mμ with dim continuum fading toward longer wave-lengths. |
| 750 | C. 5031, 2403; J. *BG*17 | Continuum from about 735–777 mμ. |

## The Isolation of Mercury Lines and Bands

The selection of filters for isolating spectral regions for visual measurements presents special problems. The ordinary spectrophotometric data available in the literature, or which can be readily obtained, are not adequate for this purpose. Such data do not pursue transmissions below about $10^{-3}$. At the ends of the visible spectrum, however, the visual sensitivity falls to much lower values than this—in our measurements to

[5] This instrument has been described by Cary and Beckman, J. Opt. Soc. Am. **31**, 682 (1941).

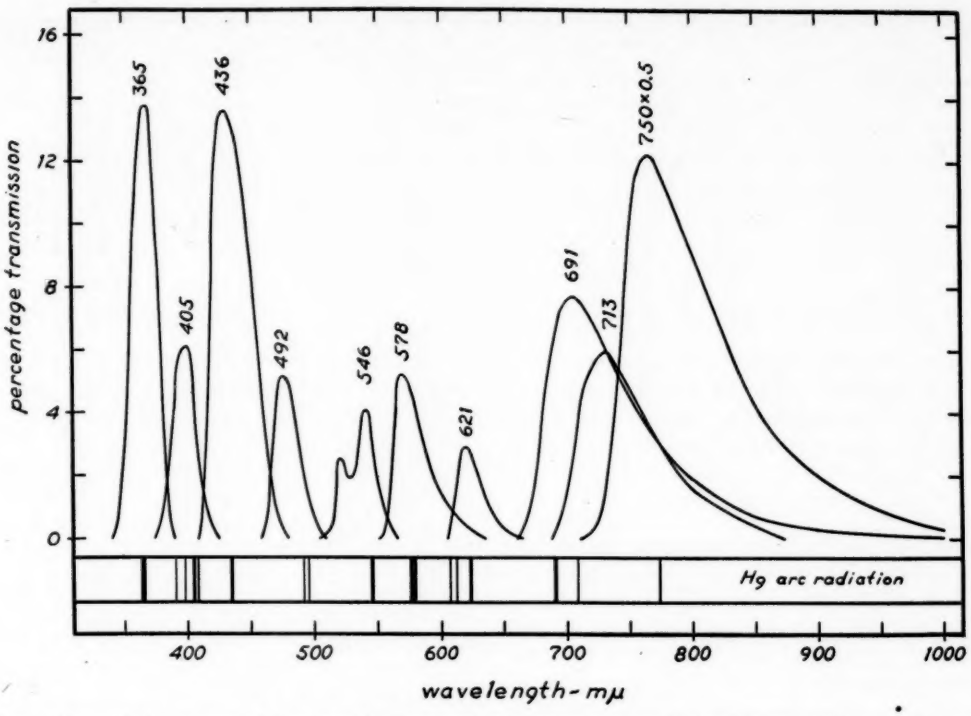



FIG. 3. Transmission spectra of filter combinations used in isolating portions of the mercury radiation. All values of percentage transmission of the 750 mμ filter have been divided by two for convenience in plotting. Only the short wave-length limbs of the transmission spectra of the 691, 713, and 750 mμ filters are effective visually; the rapidly declining sensitivity of the eye toward longer wave-lengths combines with the transmission of the filters to produce narrow bands of visible radiation, centered at the wave-lengths indicated.

less than $10^{-6}$ of the sensitivity at the maximum. Filters used for isolating these regions, therefore, may be rendered completely useless by radiation leaks toward the center of the visible, involving transmissions well below $10^{-3}$. A further complication is introduced by enormous differences in the intensities of mercury lines. A filter for isolating the feeble radiation at 405 mμ must have a transmission well below $10^{-3}$ at 436 mμ to eliminate the relatively intense line at this wave-length, to which the eye is also much more sensitive.

One can select and test filters adequately for visual measurements, therefore, only by the use of special procedures. These are discussed in an Appendix to the present paper. With their aid a satisfactory group of filter combinations has been assembled. Their constitution is listed in Table I, together with the portions of the mercury radiation which can be seen through them spectroscopically. Their transmissions, measured with the Beckman spectrophotometer, are shown in Fig. 3.

The high pressure mercury arc yields an appreciable background of continuous radiation in addition to its characteristic lines. The background, however, does not rival the lines in brightness except in the region about 610–670 mμ, and in the far red. It constitutes, therefore, a substantial portion of the radiation transmitted by the filters only at 621 and 750 mμ. In all cases a wave-length "center of luminosity" of the total transmitted light—continuum and lines—has been estimated by repeated settings of the spectroscope cross-hairs; this is the wave-length used to characterize the filter transmissions.

The broad transmission of the filters employed in the red, shown in Fig. 3, is not effective visually. The rapidly decreasing sensitivity of the eye cuts the luminosity of these radiations

TABLE II. Stability of radiation of the high pressure mercury arc. Energies at various wave-lengths, measured at intervals over a period of 17 months, at the exit field of the adaptometer. Energies are in arbitrary units, relative to that at 578 mμ. Only data involving filter combinations used continuously for long periods are shown; omissions in the table mark points at which filter combinations were changed.

| Radiation (mμ) | Relative energy | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | 6–25–43 | 8–3–43 | 11–11–43 | 1–4–44 | 4–18–44 | 4–24–44 | 7–23–44 | 11–29–44 | Averages |
| 365 | — | — | 10.12 | 10.13 | 9.70 | 9.60 | 10.21 | 9.61 | 9.90 |
| 405 | 1.21 | 1.24 | 1.18 | 1.20 | 1.21 | 1.21 | 1.23 | 1.17 | 1.21 |
| 436 | 6.02 | 5.96 | 5.66 | 5.96 | 6.31 | 6.27 | 6.25 | 6.66 | 6.14 |
| 546 | — | — | — | 1.10 | 1.08 | 1.08 | 1.10 | 1.16 | 1.10 |
| 578 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| 621 | — | 0.060 | 0.062 | 0.063 | 0.067 | 0.067 | — | — | 0.064 |
| 750 | — | — | 1.95 | 2.03 | 1.95 | 1.90 | 1.83 | 2.18 | 1.97 |

sharply at the long wave-length side of the transmission maximum, and the bands of radiation visible in the spectroscope through these filters are therefore very narrow.

### Energy Calibration

The basic devices for measuring the energy content of radiation of various wave-lengths are delicate and cumbersome. The portability of the present adaptometer has greatly contributed to its usefulness. It was important that the calibrating arrangements be equally portable so that relative energies could be determined at any time. This was particularly desirable since no information was available originally concerning the stability of the mercury arc radiation during long periods of use.

For these reasons a convenient secondary means of energy calibration has been employed, a photo-cell of the barrier layer type (the Weston Photronic Cell, Model 594, Type 1). The spectral response of this cell in turn has been calibrated relative to that of a thermopile.[6] At the test field of the adaptometer this cell generates at all wave-lengths sufficient current to be read directly with a table-type microammeter or a small enclosed-scale galvanometer.

In this way the relative energy content of each of the radiations transmitted by the various filters has been measured at a single wedge setting. This has been done periodically. After the adaptometer had been in use for several months it emerged that the relative energy calibration had remained unchanged, and this has since been maintained. The stability of the lamp in this regard is extraordinary. A series of relative energy measurements made over a period of seventeen months of regular use is shown in Table II. The reproducibility of the measurements is very satisfactory, and no systematic changes whatever are perceptible.

Such measurements provide factors with which the wedge settings can be converted to energy values. The dial readings from the adaptometer are first evaluated in terms of wedge calibration, and are then converted to relative energies by applying the energy factor for the specific filter employed.

No attempt has been made as yet to evaluate the absolute energy content of the flashes delivered by the adaptometer. This should be difficult in any case because of peculiarities of the present arrangements for producing flashes. Ideally the shutter used for this purpose should transmit a flash of square wave form—one in which the light goes on and off instantaneously. The photographic shutter yields instead something resembling a sine-wave. In the present instrument also the shutter cuts a light path ovoid in cross-section. These are not insurmountable difficulties, but up to the present we have not attempted to cope with them, largely because in the investigations for which the instrument has been used relative energies are virtually as informative as absolute values.

## APPENDIX

### Radiation Leakage in Color Filters

#### *The Problem*

Ordinary spectrophotometric data are of very limited use in selecting color filters for visual measurements. Such data as are available in the commercial literature or as

[6] This calibration was performed at the Electrical Testing Laboratories of New York City.

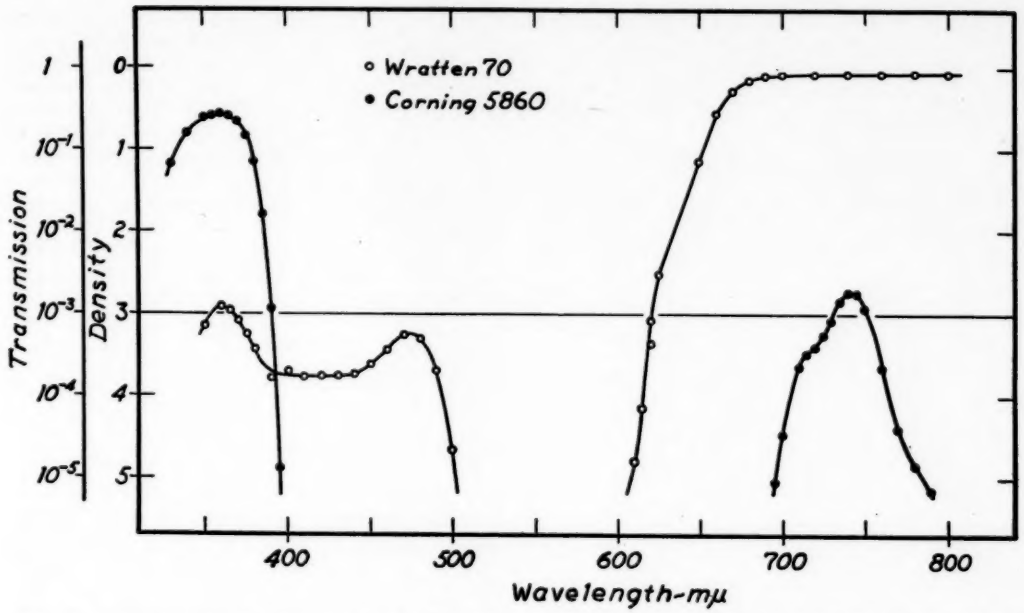



FIG. 4. Spectrophotometric transmissions of the red Wratten *70* filter, and of the Corning *5860* ("Violet Ultra") filter. An abscissa drawn at transmission $10^{-3}$ indicates the lower limit of ordinary spectrophotometric measurement. Low transmission spectrophotometery reveals radiation leaks in both filters which, though involving almost entirely transmissions below $10^{-3}$, are highly effective visually.

can be readily obtained in the laboratory, usually do not pursue spectral transmissions below about $10^{-3}$. At the ends of the visible spectrum, however, the visual sensitivity falls to much lower values than this—in our measurements to less than $10^{-6}$ of the sensitivity at the maixmum. A radiation leak near the visibility maximum therefore might involve such low transmissions as not to appear at all spectrophotometrically, and still might be much more effective visually than wave-lengths in which the filter transmission is high.

This situation is further complicated by characteristic differences in the spectra of various sources. The tungsten filament radiation is relatively rich in the red, and so throws a special burden upon filters used for isolating the violet. The mercury radiation is very rich in the violet and ultraviolet, and here it is particularly difficult to isolate the red. Filters for isolating portions of the mercury arc spectrum must cope also with enormous differences in the intensity of mercury lines.

A few examples will illustrate the nature of the problem. Wratten filters *70*, *89* and *89A*, all of which appear from spectrophotometric data to be adequate for isolating the far red, transmit enough of the mercury radiation at lower wave-lengths to invalidate them for this purpose. Conversely, the Corning glass *5860*, used for isolating the mercury radiation at 365–366 mμ, passes enough visible red to cause a considerable error in visual measurements. So also the Corning combination, *3385* and *5543*, which appears spectrophotometrically to be appropriate for isolating the weak mercury radiation at about 492 mμ, actually transmits the very intense green line at 546 mμ much more brightly, and even the yellow lines at 577 and 579 mμ as brightly as it does the 492 mμ radiation.

### *Low Transmission Spectrophotometry*

To a degree this situation can be met by extending the range of spectrophotometry to transmissions below $10^{-3}$ by the use of special procedures.

(a) In certain instances filters can be obtained of the same composition but of lower density than those to be employed—thinner filters, if of glass; or if of some dyed vehicle, filters in which the dye is more dilute. The transmissions of such relatively transparent filters can be measured at wave-lengths at which the original filter is too dense for direct measurement; and by computation, if Beer's and Lambert's Laws are applicable, the transmission of the denser filter can be estimated.

An example of results obtained in this way is shown in Fig. 4. In this case the Corning filter *5860* was first measured in its standard polished thickness of 5.0 mm between 320 and 385 mμ. Then a second filter from the same melt, but ground to a thickness of 1.50 mm was measured throughout this region and the remainder of the spectrum. The density of the thinner filter, multiplied by the ratio of thicknesses, yielded the density of the standard filter in regions in which its direct measurement was impossible. It emerged that the standard *5860* filter, recommended for

TABLE III. Test of filters by reinforcement. In this test the absolute threshold of vision is measured through a filter supposed to isolate a restricted region of the spectrum, and again through this filter combined with another of similar properties. Red filters are thus tested by reinforcement with the red Corning 2412, and an ultraviolet filter is tested with the Jena UG-1. If the filter being tested transmits only the expected wave-lengths, the added filter has the effect calculated from its spectrophotometric transmission at these wave-lengths. If the test filter, however, has a radiation leak elsewhere in the spectrum, the added filter—with very low transmission at these supposedly excluded wave-lengths—has an exaggerated visual effect.

| Filter | Composition | Change in log visual threshold due to added filter: Calculated | Observed |
|---|---|---|---|
| Red (691 mμ) | W. 91, C. 3962 | | |
| This +C. 2412 | | 0.05 | 0.08 |
| Red (750 mμ) | C. 2403, C. 5031 | | |
| This +C. 2412 | | 0.06 | 0.03 |
| Ultraviolet (365 mμ) | C. 5860, C. 738, J. BG-12 | | |
| This +J. UG-1 | | 0.22 | 0.16 |
| Red (690 mμ) | W. 89, C. 3962 | | |
| This +C. 2412 | | 0.05 | 0.31 |
| Red (680 mμ) | W. 70, J. *BG*-19 | | |
| This +C. 2412 | | 0.05 | 1.44 |
| Red (>640 mμ) | W. 70 | | |
| This +C. 2412 | | 0.06 | 1.32 |

Note: C. = Corning, J. = Jena, and W. = Wratten.

isolating near ultraviolet radiation, possesses a low transmission band in the red, maximal at about 742 mμ. Though this involves almost entirely transmissions lower than can be detected by direct spectrophotometry, it is sufficiently effective visually to introduce a considerable error.

(*b*) It is possible also to extend the range of spectrophotometry by measuring the transmission of a filter, first relative to a true blank, then relative to other filters of known density. Relative to air, for example, transmissions can be pursued to $10^{-3}$; then relative to a filter which itself transmits perhaps $10^{-3}$ of the incident light, the transmission of the original filter can again be measured down to $10^{-3}$ of this, or $10^{-6}$. In this way the range of measurement can be extended downward to the limit of sensitivity of the measuring instrument.

The transmission spectrum of the red Wratten *70* filter, measured by this means, is shown in Fig. 4. The transmission of the *70* filter was first measured relative to air between 625 and 800 mμ. It was then remeasured relative to a "neutral" filter of known transmission (approximately 1 percent) between 350 and 400 mμ. The transmission of the *70* relative to the neutral filter, multiplied by the transmission of the neutral filter relative to air, yielded the transmission of the *70* filter relative to air in this spectral region. A third series of measurements of the *70* filter relative to a still denser neutral filter (transmissions 0.1–0.7 percent) yielded data involving still lower transmissions. It emerged that in addition to its proper transmission in the red, the Wratten *70* filter possesses transmission bands maximal at about 360 and 470 mμ, with a broad region of low transmission lying between them. Though virtually all of this transmission is lower than can be detected by spectrophotometry directly, it is sufficient to invalidate the *70* filter, as is shown below, for visual measurements.

Such methods as these, even when available and adequate to the purpose, entail an almost prohibitive amount of special effort. They mean, in effect, extending the spectrophotometry of filters considered for use in visual measurements to transmissions of $10^{-5}$ and lower. Without some independent indication of the existence of a spectral leak and its approximate location, these measurements should have to be closely spaced throughout the visible spectrum. Such information does not now exist, and its preparation represents a formidable undertaking.

The evaluation of such information visually represents a further problem. The visual effectiveness of a radiation leak depends not only on the transmission of the filter, but on the energy distribution of the source, and on the spectral sensitivity of the eye. Thus the low wave-length leakage of the Wratten *70* filter is much more disturbing with the mercury arc, which is very rich in violet and near ultraviolet radiation, than with the tungsten filament; and is much more disturbing in the scotopic eye, which is relatively sensitive to low wave-lengths, than in the photopic eye. The red leakage of the Corning *5860* filter exhibits just the reverse relations: it is most effective with the tungsten radiation and in the photopic eye.

For all these reasons, simple and direct *visual* tests for radiation leaks in color filters are highly desirable. Two such tests are described below, referred to respectively as tests by reinforcement and by exclusion.

TABLE IV. Test of filters by exclusion. If a filter which supposedly transmits radiation in a restricted region is combined with one which excludes this region though transmitting elsewhere, the combination should be opaque, or may transmit a small fraction of radiation in the region of overlapping transmissions. Here the Jena *BG-18*, which excludes visible radiation longer than about 650 mμ, is used to test red filters; the combinations should be opaque or may transmit some visible red from an intense source. The Corning *3060* is similarly used to test an ultraviolet filter; the combination should be opaque.

| Filter | Combined with | Appearance of sun through combination |
|---|---|---|
| Wratten 70 | Jena BG-18 | Bright violet |
| Wratten 87 | Jena BG-18 | Very deep red |
| Wratten 88 | Jena BG-18 | Opaque |
| Wratten 88A | Jena BG-18 | Opaque |
| Wratten 89 | Jena BG-18 | Deep Orange |
| Wratten 89A | Jena BG-18 | Bright blue-green |
| Jena RG-5 | Jena BG-18 | Deep red |
| Jena RG-8 | Jena BG-18 | Deep red |
| Corning 5860 | Corning 3060 | Deep red |

### *Reinforcement*

In the test by reinforcement one measures the intensity required to produce some specific visual effect—for example, a minimal stimulus, or a specific brightness—through the filter in question. Then one reinforces this filter with a second filter having much the same transmission properties, and repeats the visual measurement. For example, if one is testing a filter supposed to isolate the far red, one makes a visual measurement first with this alone, and then with this combined with a second red filter. If the filter being tested has no low wave-length leakage, the second filter superimposed on it will have only the effect calculated from its spectrophotometric transmission. If, however, the test filter does have a radiation leak at low wave-lengths, the second filter, with its very low transmission at these wave-lengths, will have a greatly exaggerated effect upon the visual result.

Applications of this test are shown in Table III. The absolute threshold of vision was measured with filters supposed to isolate the red or the ultraviolet, and again with these reinforced with other filters of similar properties.

The red combination, Wratten *91* and Corning *3962*, used above for isolating the 691 mμ radiation, was tested alone and reinforced with the red Corning *2412* filter. The change in log threshold expected from the transmission of the Corning *2412* at 691 mμ is 0.05; the change observed is 0.08. This agreement is satisfactory. It shows that the Wratten *91* filter has no appreciable radiation leakage at low wave-lengths. The red combination, Corning *2403* and *5031*, used above for isolating the 750 mμ radiation, also passes this test satisfactorily.

Similarly the combination used for isolating the 365 mμ radiation, tested by reinforcement with the Jena *UG-1* filter, gives results in good agreement with expectation.

On the other hand, the reinforcement with the Corning *2412* of a red combination based on the Wratten *89* filter has an exaggerated effect. The rise in log threshold expected from the spectrophotometric transmission is 0.05, while that observed is 0.31. Short wave-length leakage by this filter produces a discrepancy in the threshold measurement of about 0.26 log unit.

The Wratten *70* filter displays a much more striking discrepancy. Its reinforcement with the Corning *2412*, calculated to raise the threshold 0.05 log unit, actually raises it about 1.3 log units. In this case, radiations of lower wave-length are about 20 times more effective than the red transmission in stimulating vision.

By this test, therefore, the first three filters shown in Table III are demonstrated to be effective in isolating the ends of the visible spectrum, while the Wratten *89* and *70* filters are shown to be unsuitable for this purpose.

### *Exclusion*

In the test of filters by *exclusion* radiation leakage may be viewed directly. In this test the filter in question is combined with another with inverse properties—one which is opaque at wave-lengths at which the transmission of the test filter is high, and is transparent where the test filter supposedly is opaque. On looking through such a combination at an intense source one should find it to be opaque, or at most to transmit a small amount of light in the region of overlapping transmission of the filter employed. If, however, the filter being tested transmits in other regions than those shown spectrophotometrically, this transmission—since it is passed by the added filter—is immediately apparent. Tests of this type are shown in Table IV.

For testing red filters in this way I have used the Jena *BG-18*, which excludes visible radiation longer than about 650 mμ, though its transmission is high throughout the rest of the spectrum. The Wratten *70*, combined with this filter, is seen to transmit sunlight as bright violet. Similarly, the Wratten *89A* filter, tested in this way, transmits blue-green; the Wratten *89* transmits orange. These are the radiation leaks which invalidate these red filters for visual measurements (cf. Table III). On the other hand, the red Wratten filters *87*, *88*, and *88A*, and the red Jena filters *RG-5* and *RG-8*, all pass this test efficiently.

The Corning *5860* filter, used to isolate the mercury line at 365 mμ, is stated to transmit no visible red. If this is combined, however, with the Corning *3060*, which excludes the ultraviolet while transmitting the red, the combination is seen to transmit a deep red radiation. For this reason the Corning *5860* filter must be combined with a filter which eliminates the far red in order to be used for visual measurements. We have employed for this purpose the Jena *BG-12*, as shown in Table I above.

The most thorough application of the exclusion method for testing filters consists in their inspection with a good spectroscope. To make this effective a very intense source of light must be employed. We have used the General Electric H-4 mercury lamp, with the image of its capillary discharge focused directly on the entrance slit of the spectroscope.

With the Wratten *70* filter in the light path, one sees in the spectroscope—in addition to the red radiation which it should transmit—dim violet lines at 405 and 408 mμ, the bright violet line at 436 mμ, and the blue-green lines at 492 and 495 mμ. Through the red Wratten *89* filter one can see the yellow lines at 577 and 579 very clearly, the green line at 546 mμ much more dimly, and the violet line at 436 mμ as just visible. Through the red Wratten *89A*, the lines at 579, 577 and 546 mμ appear clearly, there is a bright region of transmission in the region of the lines at 492 and 495 mμ, and the 436 mμ line is just visible. On the other hand, with an efficient red filter such as the Jena *RG-8*, no radiation whatever is visible below the transmission in the red, even though the slit of the spectroscope is opened widely.

**SUMMARY**

A spectral adaptometer is described which uses as source a high pressure mercury arc. From its radiation ten spectral regions, centered at 365, 405, 436, 492, 546, 578, 621, 691, 713, and 750 mμ, are isolated with color filters. The in-

tensity at each of these wave-lengths is varied with an annular photographic wedge. A variable fixation point permits the image of the test field to be located centrally or at any point within a radius of 12° of the fovea. A camera shutter exposes the test field for flashes of graded duration. A sliding diaphragm occludes either eye at will. The entire apparatus is portable.

In selecting filter combinations for rigorously isolating portions of the mercury radiation, ordinary spectrophotometric data were found to provide an inadequate guide. Radiation leaks in color filters which involve transmissions below $10^{-3}$, though not usually measured spectrophotometrically, may be very effective visually. This situation is discussed in an Appendix, with the application to it of two methods of low-transmission spectrophotometry and two types of direct visual test for radiation leakage in color filters.

JOURNAL OF THE OPTICAL SOCIETY OF AMERICA VOLUME 35, NUMBER 3 MARCH, 1945

# The Geometrical Solutions of Some Color Mixture Problems

M. M. GOUREVICH
*Photometrical Laboratory, State Optical Institute, U.S.S.R.*

(Received December 15, 1944)

1. THE methods of solving color mixture problems have been discussed repeatedly in various articles.[1–8]

At present the following fundamental procedures may be considered as established, though they seldom are clearly expressed.

Three primary colors $R$, $G$, $V$ are chosen in such a way that no one of them can be produced by mixing the other two; they may be represented by three vectors drawn in various directions from the origin point $O$ (Fig. 1). Through the end points of these three vectors a plane $P$ is drawn. We shall consider each color represented by any point $C$ of the plane $P$ as a unit color and the vector $OC$ as a unit vector in the $OC$ direction.

The position of the point $C$ on plane $P$ depends upon the quality of the color and determines its chromaticity.

The ratio of the vector length $OQ$ to the length of the corresponding unit vector $OC$ we shall call the module of color $Q$.*

By using the notion of the module of a color we are able to formulate the following statements useful in color computations:

(a) The module of a mixture of several colors is equal to the sum of modules of the mixed constituents.

(b) For the purpose of finding the position of the point $C$ representing the chromaticity of the mixture of several colors the rules for the determination of the center of gravity may be used. When defining the center of gravity the masses of colors are to be considered as equal to their modules and the positions of these masses as coinciding with the positions of the points representing the chromaticities of the mixed constituents on the plane $P$.

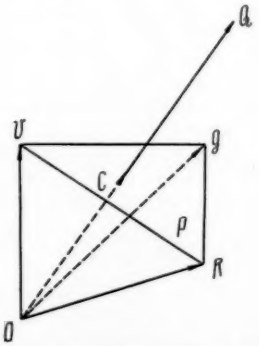



FIG. 1. The module of a color is the ratio of a vector length $OQ$ to the corresponding unit vector $OC$.

[1] H. E. Ives, J. Frank. Inst. **180**, 673 (1915); **195**, 23 (1923).
[2] C. L. Froelich, J. Opt. Soc. Am. **9**, 31 (1924).
[3] I. G. Priest, J. Opt. Soc. Am. **9**, 503 (1924).
[4] J. Guild, Trans. Opt. Soc. **26**, 95–108, 139–174 (1924–25).
[5] D. L. MacAdam, J. Opt. Soc. Am. **28**, 103–111 (1938).
[6] N. G. Boldireff, Trans. State Opt. Inst. No. III, p. 25 (1939).
[7] M. M. Gourevich, J. Tech. Phys. **9**, 1039 (1939).
[8] S. O. Maisel and E. S. Ratner, Trans. WEI, No. 56 (1941).

* It may be called also "quantity of color" or "weight of color." It is obvious that the module of color depends upon the primary colors, i.e., upon the color system.

Each color $Q$ may be considered as a mixture of three primaries $R$, $G$, $V$ the quantities $r'$, $g'$, $v'$ of which represent the coordinates of the color $Q$ in this system.

$$Q=r'R+g'G+v'V. \tag{1}$$

By using statement (a) we see that the module $m$ of the color $Q$ is

$$m=r'+g'+v', \tag{2}$$

since $r'$, $g'$, $v'$ may be regarded as the modules of colors $r'R$, $g'G$, and $v'V$.

Thus the third statement comes out:

(c) The module of a color is equal to the sum of its coordinates.

Hence it follows, that the modules of all unit colors are equal to a unity, since the equation of the plane $P$ is

$$r+g+v=1.*$$

At the same time the module of a color $Q$ may be expressed as a ratio of each of its coordinates to the corresponding coordinate of the unit color $C$;

$$m=r'/r=g'/g=v'/v.$$

2. It is well known that, in the case of a mixture of two colors $Q_1$ and $Q_2$ having the chromaticities $C_1$ and $C_2$ and modules $m_1$ and $m_2$, the point $C$ representing the chromaticity of the mixture is situated on the line $C_1C_2$ and that

$$CC_1:CC_2=m_2:m_1. \tag{3}$$

In the case of a mixture of three colors $Q_1$, $Q_2$, and $Q_3$ having chromaticities $C_1$, $C_2$, and $C_3$ (Fig. 2), the position of the point $C$, which depicts the chromaticity of their mixture, is connected to the positions of points $C_1$, $C_2$, $C_3$ by simple relations. If, as before, $m_1$, $m_2$, $m_3$ and $m$ are the modules of the three constituents and that of their mixture correspondingly, then

$$m=m_1+m_2+m_3, \tag{4}$$

and

$$\begin{aligned} CA_1/C_1A_1&=m_1/m,\\ CA_2/C_2A_2&=m_2/m,\\ CA_3/C_3A_3&=m_3/m. \end{aligned} \tag{5}$$

It is evident that

$$CA_1/C_1A_1+CA_2/C_2A_2+CA_3/C_3A_3=1.$$

* The coordinates of unit colors in the I.C.I. color chart are usually denoted by $x$, $y$, $z$ instead of $r$, $g$, $v$.

These expressions are useful for the geometrical solutions of color mixture problems.

3. The interrelations between the modules of colors and their chromaticities, allowing the use of geometrical methods of computation, do not indicate immediately the luminance of colors which is in many cases most important. Among the articles published, as far as the author knows, only one[6] gives a simple method of geometrical determination of luminance ratios. There exist,

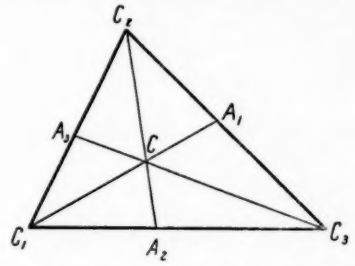



FIG. 2. The ratios $CA_1/C_1A_1$, $CA_2/C_2A_2$, $CA_3/C_3A_3$ are equal to the ratios of the modules of the mixed colors to the module of their mixture.

however, some other means of getting analogous results.

It is evident that the I.C.I. color system is especially suitable for determining the luminance ratios. The foregoing relationships will be applied to it here. It is also well known that in the I.C.I. color system the second color coordinate $y'$ is equal to the luminance $B$ of the color, i.e., $y'=B$.

The module of a color in the I.C.I. color system may be formulated

$$m=x'+y'+z'=x'/x=y'/y=z'/z=B/y. \tag{6}$$

If we consider once more the case of the mixture of two colors $Q_1$ and $Q_2$ (Fig. 3) then according to Eqs. (3) and (6) a relation

$$CC_1:CC_2=m_2:m_1=B_2/y_2:B_1/y_1$$

will be found, where $B_1$ and $B_2$ are the luminances of $Q_1$ and $Q_2$ and $y_1$ and $y_2$ are the ordinates of their chromaticities $C_1$ and $C_2$.

After the line $C_1H_2$ is drawn and the point $N$ of its intersection with $CH$ is found it becomes clear that

$$B_2/B_1=CN/NH.$$

Indeed, from Fig. 3 we see that

$$CN/y_2=CC_1/C_1C_2 \quad \text{and} \quad NH/y_1=CC_2/C_1C_2,$$

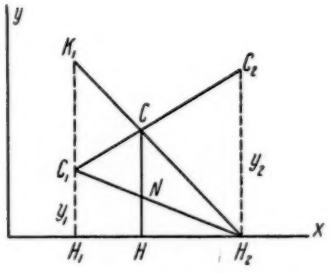



FIG. 3. On the I.C.I. color-chart the ratio $CN/NH = B_2/B_1$, $B_1$ and $B_2$ being the luminances of two mixed colors having chromaticities $C_1$ and $C_2$.

whence

$$CN/NH = CC_1/y_1 : CC_2/y_2 = B_2/B_1.$$

If $C_1$ represents the chromaticity of a white and $C_2$ that of a spectrum color, then the colorimetric purity $p$ of their mixture is

$$p = CN/CH.$$

If we examine the converse problem and seek the position of the point $C$, the ratio of the luminances $B_1$ and $B_2$ being given, then we have to proceed as follows. Producing the line $H_1C_1$ we plot on it the point $K_1$, so that

$$C_1K_1 = y_1(B_2/B_1)$$

and join $K_1$ to $H_2$. It is obvious that $C$—the point of its intersection with $C_1C_2$—represents the chromaticity of the mixture.

In mixing three colors $Q_1$, $Q_2$, $Q_3$ having chromaticities $C_1$, $C_2$, $C_3$ (Fig. 4) it is often interesting to find out which part of the total luminance ($B$) of the mixture constitutes the luminance of any of its components as, e.g., the luminance $B_1$ of the color $Q_1$. In other words, the ratio $B_1/B$ is to be found. For the solution of this problem the line $C_1C$ must be produced to meet $C_2C_3$ at $A_1$. Then after having joined $A_1$ to $H_1$ and denoted the point of its intersection with $CH$ by $N$, we can see that

$$CN/y_1 = CA_1/C_1A_1.$$

Furthermore, Eq. (5) shows that

$$CA_1/C_1A_1 = m_1/m = B_1/y_1 : B/y,$$

whence

$$B_1/B = CN/y = CN/CH,$$

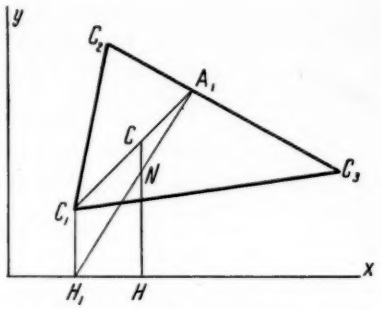



FIG. 4. If $C$ represents the chromaticity of a mixture of three constituents then $CN/CH = B_1/B$, $B_1$ being the luminance of one constituent ($C_1$) and $B$ that of the mixture.

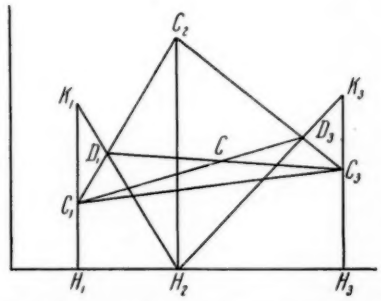



FIG. 5. This scheme demonstrates the method of finding the chromaticity $C$ of a mixture of three constituents (the chromaticities of constituents being $C_1$, $C_2$, $C_3$) when the luminance ratios $B_1 : B_2 : B_3$ are given. The ratio $C_1K_1/C_1H_1 = B_2/B_1$ and correspondingly $C_3K_3/C_3H_3 = B_2/B_3$.

where $B_1$ and $B$ represent the luminances of the color $Q_1$ and of the mixture, respectively.

Finally, let us suppose that we know the ratios of the luminances $B_1 : B_2 : B_3$ of three mixed colors having chromaticities $C_1$, $C_2$, $C_3$ (Fig. 5) and that we have to find the chromaticity $C$ of their mixture. In this case we have to proceed as follows.

Having produced the lines $C_1H_1$ and $C_3H_3$, points $K_1$ and $K_3$ are to be marked on them so that

$$C_1K_1 = y_1(B_2/B_1) \quad \text{and} \quad C_3K_3 = y_3(B_2/B_3).$$

Points $K_1$ and $K_3$ are joined to $H_2$ and the points $D_1$ and $D_3$ of intersection of $K_1H_2$ and $K_3H_2$ with $C_1C_2$ and $C_2C_3$ must be joined to $C_3$ and $C_1$, respectively. The point of intersection of $D_1C_3$ and $D_3C_1$ then determines the chromaticity $C$ of the mixture which is sought.

JOURNAL OF THE OPTICAL SOCIETY OF AMERICA VOLUME 35, NUMBER 3 MARCH, 1945

# Standard Response Functions for Protanopic and Deuteranopic Vision

DEANE B. JUDD
*National Bureau of Standards, Washington, D. C.*
(Received September 11, 1944)

## I. INTRODUCTION

WHEN the two halves of a photometric field are illuminated to the same degree with light of the same spectral composition, they cannot be distinguished; that is, they produce a color match. But it is possible also to produce a color match between lights of different spectral compositions; such lights are called metamers,[22, 51] and are said to form metameric pairs. The functions which give the conditions to be met by the spectral composition of any two lights in order that they shall form a metameric pair for a given observer serve to define the characteristics of that observer; and in 1931 the International Commission on Illumination adopted such functions which define the standard observer.[23, 27]

Protanopic and deuteranopic vision are the two most common forms of dichromatic vision, each form occurring in about one percent of otherwise normal human males.[31] They are known as reduction forms because neither a protanope nor a deuteranope ordinarily distinguishes the members of any pair of normal metamers; that is, a color match set up by a normal observer will not be objected to by a dichromat unless the pigmentation of his eye media, including the macula, be grossly different from that of the normal observer. The standard color specifications are therefore generally valid for dichromats in the sense that if two lights have the same standard color specifications, no average dichromat will be able to distinguish them.

These specifications are, however, unnecessarily complicated for such observers. In addition to normal metamers, dichromats find many other pairs which they can not distinguish. These additional metamers are often called confusion colors; they are colors which are distinct to the normal although identical to the dichromat. The normal color specification consists of three numbers, because the normal visual system is capable of three independent modes of variation and on that account is often called trichromatic. The dichromatic specification need consist of but two numbers. These two numbers may be derived from the standard color specification in a simple way. The connection between trichromatic and dichromatic specifications was indicated in complete detail by Maxwell in 1855[44] by appeal to the Young theory of vision. The original simple form of this theory has been disproved, the extensions of it vary with the phenomena to be explained, and there is doubt whether the theory, however extended, can yield more than a partial explanation of the complicated facts of vision. However, the connection between dichromatic and trichromatic color specifications first worked out by Maxwell on theoretical ground does not depend upon the portions of Young's hypothesis now given up; on the contrary this connection has been repeatedly proven to be correct. It is the purpose of this paper to derive, by this principle, response functions for protanopic and deuteranopic vision, to show how they may be used to find whether or not two colors will be confused by an average protanope or average deuteranope, and finally to point out a few of the theoretical implications of the functions chosen.

## II. THEORY

Maxwell wrote in his letter of January 4, 1855, to Dr. George Wilson:[44] "If we find two combinations of colours which appear identical to a Colour-Blind person, and mark their positions on the triangle of colours (now called the Maxwell triangle), then the straight line passing through these points will pass through all points corresponding to other colours, which, to such a person, appear identical with the first two.

"We may in the same way find other lines passing through the series of colours which appear alike to the Colour-Blind. All these lines either pass through one point or are parallel, according to the standard colours which we have

assumed, . . . Knowing this law of Colour-Blind vision, we may predict any number of equations which will be true for eyes having this defect.

"The mathematical expression of the difference between Colour-Blind and ordinary vision is that colour to the former is a function of two independent variables, but to an ordinary eye, of three; and that the relation of the two kinds of vision is not arbitrary, but indicates the absence of a determinate sensation, depending perhaps upon some undiscovered structure or organic arrangement, which forms one-third of the apparatus by which we receive sensations of colour."

The straight lines on the Maxwell triangle serve to indicate the chromaticity confusions of the color-blind observer. The point of intersection of these lines indicates the normal chromaticity of the primary process not possessed by the color-blind. Suppose for the protanope that the chromaticity-confusion lines are copunctal at $(x_p, y_p)$ on the $(x, y)$-plane of the standard I.C.I. colorimetric coordinate system,[23,27] and suppose for the deuteranope that the corresponding point is $(x_d, y_d)$. Let us inquire how to derive from the functions (of wave-length) $(X, Y, Z)$ defining the standard normal observer, three new functions $(W_d, W_p, K)$ such that all three taken together represent the normal observer, such that $(W_d, K)$ taken together represent average deuteranopic vision, and such that $(W_p, K)$ taken together represent average protanopic vision. This terminology $(W_d, W_p, K)$ is taken from v. Kries* who followed König[37] closely. The symbol $W$ is intended to suggest "warm;" and $K$, "cold" (kalt), corresponding to whatever warm color (orange, yellow, greenish yellow) and whatever cold color (blue or violet) is sensed by red-green-blind observers.

The derivation consists of making new choices of primary processes such that two of them correspond to $(x_p, y_p)$ and $(x_d, y_d)$, respectively. It was pointed out by König in 1886[37] for this very purpose, and by many others since,[13,24,25,43,54] for other purposes, that such choices result from defining the new functions $(W_d, W_p, K)$ as weighted sums of the old, thus:

$$\begin{aligned} W_d &= k_1X + k_2Y + k_3Z, \\ W_p &= k_4X + k_5Y + k_6Z, \\ K &= k_7X + k_8Y + k_9Z, \end{aligned} \tag{1}$$

where $k_1$ to $k_9$ are constants which may be given any arbitrary values whose determinant differs from zero:

$$\begin{vmatrix} k_1, & k_2, & k_3 \\ k_4, & k_5, & k_6 \\ k_7, & k_8, & k_9 \end{vmatrix} \neq 0. \tag{1a}$$

Since the point $(x_p, y_p)$ corresponds to the primary $W_d$ not possessed by the protanope, $W_p = K = 0$, for $X/(X+Y+Z) = x_p$, and $Y/(X+Y+Z) = y_p$, and we may write from Eq. (1):

$$\begin{aligned} k_4x_p + k_5y_p + k_6z_p &= 0, \\ k_7x_p + k_8y_p + k_9z_p &= 0, \end{aligned} \tag{2a}$$

where $z$ is defined as $Z/(X+Y+Z)$ and is equal to $1-x-y$.

Similarly, since the point $(x_d, y_d)$ corresponds to the primary not possessed by deuteranopes we may write:

$$\begin{aligned} k_1x_d + k_2y_d + k_3z_d &= 0, \\ k_7x_d + k_8y_d + k_9z_d &= 0. \end{aligned} \tag{2b}$$

The four conditions expressed by Eqs. (2a) and (2b) are the only conditions that have to be met to insure that $W_p$ and $W_d$ represent the primary color processes not possessed by the deuteranope and protanope, respectively. Since nine conditions are required to determine the nine constants of Eq. (1), it may be seen that many coordinate systems can serve both for normal trichromatic visual systems and for dichromatic systems by neglect of the one or the other of two of the three normal components.

Two of the remaining five degrees of freedom are required to insure, as is convenient, that the equal-energy stimulus be kept as the basic stimulus[14] of the system, that is, that it be represented at the center of the Maxwell triangle of the new coordinate system as well as in the standard I.C.I. system. This requirement is satisfied if the distribution curves of the new

* Reference 39, p. 164.

primary color processes throughout the equal-energy spectrum be adjusted, like those of the I.C.I. system, to equal areas; that is, if $k_1+k_2+k_3=k_4+k_5+k_6=k_7+k_8+k_9$. Another degree of freedom must be expended to set the arbitrary units in which the distribution curves are expressed. The three conditions together may be expressed conveniently as:

$$\begin{aligned} k_1+k_2+k_3&=1,\\ k_4+k_5+k_6&=1,\\ k_7+k_8+k_9&=1. \end{aligned} \tag{3}$$

König[37] used the other two conditions to fix the third primary at an imaginary color of dominant wave-length near to that usually corresponding to unitary blue, that is, a stimulus which is perceived under ordinary observing conditions as a blue which is neither reddish nor greenish. In this way König derived the "fundamental sensation" curves incorporated by Ladd-Franklin[41] into her theory of color vision. Following further studies of dichromatic vision, however, König gave up the attempt to make any of the primaries correspond to a unitary hue.[38] In this similar reduction of present-day data on the vision of dichromats, it has likewise seemed advisable to pay no attention to the color perception ordinarily corresponding to the primaries adopted, but rather to strive for the simplest possible adequate representation of the data. Thus, it is possible to satisfy conditions (2) and (3) in such a way as to yield wave-length functions for the primaries each consisting of a curve possessing the single-peak shape of the luminosity function. To avoid functions for $W_d$ and $W_p$ having two maxima it is sufficient simply to require:

$$\begin{aligned} k_3&=-0.22k_1,\\ k_6&=-0.22k_4. \end{aligned} \tag{4}$$

These requirements prevent $W_d$ and $W_p$ from being large near 440 m$\mu$ by utilizing only enough of the $Z$ function to cancel approximately the secondary maximum of the $X$ function in that region.

### III. CHROMATICITY CONFUSION OF DICHROMATS

As first pointed out by Maxwell, the chromaticity confusions of either type of dichromat may be represented on the chromaticity diagram, or Maxwell triangle, for normal trichromatic vision by a family of straight lines passing through a single point, $(x_p, y_p)$ for the protanope, $(x_d, y_d)$ for the deuteranope. The first determination of these points was that by König and Dieterici[37] based upon two protanopes and two deuteranopes. From these points were derived the "fundamental-sensation" curves (Grund-Empfindungs-Curven) $R'$, $G'$, and $B'$; and from the approximate relation[30] between these curves and those of the present standard coordinate system for colorimetry, an estimate of the coordinates of the points may be made. From the previously evaluated transformation equations* from standard color specifications $(X, Y, Z)$ to $(R', G', B')$ may be found the reverse transformations†

$$\begin{aligned} X&=0.244R'-0.058G'+0.014B',\\ Y&=0.056R'+0.150G'-0.005B',\\ Z&=0.000R'+0.000G'+0.200B'. \end{aligned} \tag{5}$$

By setting $G'=B'=0$ in Eq. (5), we find $X=0.244R'$, $Y=0.056R'$, $Z=0$, whence:

$$\begin{aligned} x_p&=0.244R'/(0.244R'+0.056R')=0.81\\ y_p&=0.056R'/(0.244R'+0.056R')=0.19. \end{aligned}$$

Similarly by setting $R'=B'=0$, we find $x_d=-0.63$, $y_d=1.63$.

Of course, no very great dependence can be placed on these results of this pioneer work. The connection with the present standard coordinate system is uncertain not only because it is based upon but two partially dark-adapted normal observers (König and Dieterici, who incidentally differed importantly from each other), but also because the basic stimulus of the system ("sunlight reaching the earth's surface through atmosphere of highest transmission") is essentially undefined. If we assume $x=0.33$ and $y=0.34$ for this basic stimulus[46] the protanopic and deuteranopic "neutral" points in the spectrum are found graphically on the $(x, y)$ plot of the standard I.C.I. system to be 496 and 511 m$\mu$, respectively. This agrees only approximately with the wave-lengths (495 and 504 m$\mu$, respectively) read from the König-Dieterici triangle;‡

* See reference 30, Eq. (7b).
† See reference 30, Eq. (2).
‡ See reference 37, Fig. 7.

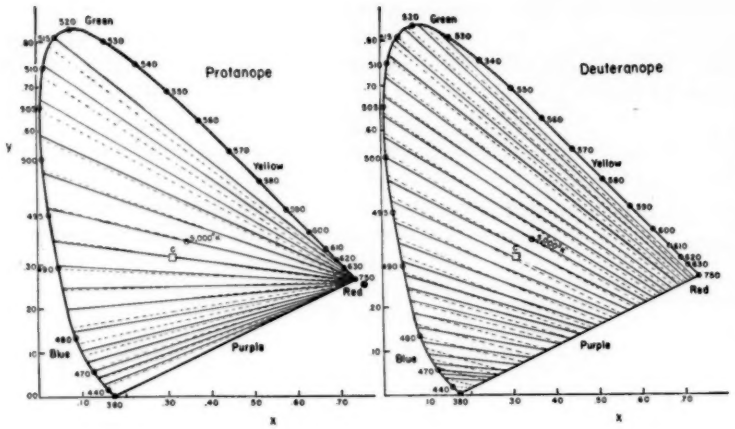



FIG. 1. Chromaticity confusions of the protanope and deuteranope (after Pitt). Note how closely the chromaticity-confusion lines (dotted) intersect at a single point for each of the two recognized types of red-green blindness. The solid lines included for comparison are copunctal (protanope at $x=0.747$, $y=0.253$; deuteranope at $x=1.000$, $y=0.000$).

and indicates that the possibilities for considerable error have been realized.

In 1935, however, Pitt[52] published data giving the average chromaticity confusions of 8 deuteranopes and 7 protanopes in terms of the *WDW* coordinate system proposed by Wright[59] for visual research because it takes account in a simple way of variations in ocular pigmentation (macula lutea, crystalline lens, humors). From the primaries of this system (spectrum lights at 460, 530, and 650 m$\mu$, respectively) and from the fact that I.C.I. standard Illuminant $B$[23,27] has in this system for the standard observer the chromaticity coordinates $r=0.249$, $g=0.399$, equations, which hold for the I.C.I. standard observer, may be derived to connect these chromaticity coordinates $(r, g)$ with those $(x, y)$ of the I.C.I. standard coordinate system:

$$x=\frac{0.874r-0.023g+0.144}{0.402r-0.222g+1.000},$$
$$y=\frac{0.354r+0.597g+0.030}{0.402r-0.222g+1.000}. \qquad (6)$$

By means of these equations the chromaticity-confusion lines shown on Pitt's Figs. 15 and 16 have been transferred to the I.C.I. system and are shown as dotted lines in Figs. 1a and 1b. It will be noted that the protanope chromaticity confusion lines plotted in Fig. 1a all run very closely through a single point ($x_p=0.747$, $y_p=0.253$) in conformity to the principle enunciated by Maxwell from the Young theory. The deuteranope chromaticity-confusion lines (Fig. 1b) are not so perfectly copunctal as this, but most of them come fairly close to the point $x_d=1.08$, $y_d=-0.08$.

In 1936, Hecht and Shlaer[15] published complete data for two dichromats, one protanope, and one deuteranope. These data include luminosity functions, wave-length discrimination, and color-mixture data. The color-mixture data take the form of the energy ratio of two spectral primaries (458.7 and 570.0 m$\mu$) required to produce a chromaticity match for each part of the spectrum. This energy ratio can be accurately defined by the dichromatic observer for the middle portion of the spectrum, but near both 460 m$\mu$ and 570 m$\mu$ a given energy ratio is found to apply to a considerable spectral range which continues to widen as the spectral extremes are approached. Table I gives average values of these energy ratios found from the Hecht-Schlaer data taking due account of the rapid decline in the ability of these observers to discriminate wave-length change near 460 and 570 m$\mu$. These data may be made independent of the degree of ocular-media pigmentation by transforming them to the red and blue primaries of the *WDW*[59] system. In this system spectral primaries (460 and 650 m$\mu$) are used, but, instead of energy units, arbitrary units such as to make the amounts of the primaries equal for 494 m$\mu$ are adopted. By the usual methods,[26] equations of transformation have been found connecting trichromatic coefficients $(r, b)$ in the *WDW*

TABLE I. Comparison of the Hecht-Shlaer color-mixture data for one protanope and one deuteranope with Pitt's data for 8 protanopes and 7 deuteranopes.

| Wave-length in mμ | Hecht-Shlaer protanope HJ | | Pitt average protanope | Hecht-Schlaer deuteranope AWG | | Pitt average deuteranope |
|---|---|---|---|---|---|---|
| | $f_p$ | $b=1-r$ | $b=1-r$ | $f_d$ | $b=1-r$ | $b=1-r$ |
| 450 | | | 1.03 | | | 1.03 |
| 460 | 150.0 | 1.00 | 1.00 | 150 | 1.00 | 1.00 |
| 470 | 6.4 | .93 | .93 | 28 | .92 | .93 |
| 480 | 2.1 | .81 | .80 | 11.2 | .79 | .80 |
| 490 | .77 | .61 | .57 | 5.0 | .62 | .57 |
| 500 | .294 | .37 | .34 | 1.5 | .32 | .34 |
| 510 | .100 | .17 | .17 | .526 | .14 | .14 |
| 520 | .033 | .06 | .09 | .182 | .05 | .06 |
| 530 | .0113 | .02 | .06 | .062 | .02 | .03 |
| 540 | .0040 | .01 | .04 | .025 | .01 | .02 |
| 550 | .0005 | .001 | .02 | .010 | .003 | .01 |
| 560 | | | .01 | .0025 | .001 | .00 |
| 570 | | | .00 | | | .00 |
| 580 | | | .00 | | | .00 |
| 590 | | | .00 | | | .00 |

system to the energy ratios $f_d$ and $f_p$ for the protanope $HJ$ and the deuteranope $AWG$:

$$r=(0.492-0.003f_p)/(0.492+0.997f_p),$$
$$b=f_p/(0.492+0.997f_p), \quad (7a)$$

$$r=(31.85-0.021f_d)/(31.85+0.979f_d),$$
$$b=f_d/(31.85+0.979f_d). \quad (7b)$$

Table I shows the color-mixture data for the protanope $HJ$ and the deuteranope $AWG$ transformed by Eqs. (7a) and (7b) to the red and blue primaries of the $WDW$ system, and for comparison it also shows the average values found by Pitt[52] for 8 protanopes and 7 deuteranopes, respectively. It is seen that the Hecht-Shlaer color-mixture data resemble those of Pitt considerably, and comparison of the discrepancies with the known wave-length discrimination of dichromats,[15, 52] and with the individual differences among the dichromats studied by Pitt* shows that the corroboration is wholly satisfactory.† It is to be concluded, therefore, that the Hecht-Schlaer data when referred to the same degree of ocular pigmentation as is included in the I.C.I. standard observer indicate the same values of $(x_p, y_p)$ and $(x_d, y_d)$ as the Pitt data.

From Fig. 1 it may be seen that a source of color temperature 5000°K ($x=0.344$, $y=0.351$) would possess for an average protanope the same chromaticity as the spectrum at 496 mμ; and for an average deuteranope, the neutral point would be at 500 mμ. As pointed out by v. Kries[38b] a dichromat having ocular-pigmentation heavier than that of the standard observer would have his neutral point shifted toward the long wave end of the spectrum; for a lightly pigmented dichromat a displacement toward the short wave

* Reference 52, p. 15.

† Although the color-mixture data for these two dichromats were not put by Hecht and Shlaer into a form suited to prediction of dichromat color matches between heterogeneous stimuli and spectrum stimuli, such a reduction has been carried out recently by Fry (reference 10). From this reduction it would be possible to determine independently the chromaticity coordinates $(x_p, y_p)$ and $(x_d, y_d)$ of the primary processes not possessed by the two forms of dichromat, instead of relying for this purpose upon Pitt's transformation from the dichromatic to trichromatic diagram of the $WDW$ system by determination of the dichromatic luminosities of the primaries. This reduction indicates, however, that the complete radiator at 5000°K should have had the chromaticity of the spectrum at 498 mμ for protanope $HJ$ and 508.5 mμ for deuteranope $AWG$, but by direct observation these wave-lengths were found to be 491.5 mμ and 495 mμ, respectively. These discrepancies amount to 5 or 10 times the just noticeable difference for these observers. Although the major part of the discrepancy for protanope $HJ$ can be eliminated by more exact methods of reducing the data and by filling in the wave-length region 380 to 470 mμ from Pitt's data, the Hecht-Shlaer data being uncertain in that region, still it would seem that no great reliance is to be placed upon the predictions of dichromat color matches from these data. These discrepancies suggest that the experimental conditions under which the luminosity functions of these observers were obtained may have brought into play a smaller retinal region than that for the color-mixture data although ostensibly the conditions were identical. Perhaps the mere fact that the luminosity and color-mixture determinations were made separately is a sufficient explanation of the discrepancy. It is doubtful therefore whether separate determination of the copunctal points $(x_p, y_p)$ and $(x_d, y_d)$ from these data would add appreciably to the information obtained by way of the $WDW$ coordinate system.

TABLE II. Individual variation in wave-length of neutral point compared to the expected influence of ocular pigmentation based upon 35 normal observers studied by Wright (reference 59).

| Source of data | Number and type of observers | Wave-length in mμ of the spectral region having the dichromatic chromaticity of a source of color temperature 5000°K. Protanope Min. | Protanope Max. | Protanope Average | Deuteranope Min. | Deuteranope Max. | Deuteranope Average |
|---|---|---|---|---|---|---|---|
| Fig. 1 | 35 normals | 489 | 504 | 496 | 494 | 508 | 500 |
| König (1884) | 7 protanopes<br>7 deuteranopes | 495 | 499 | 497 | 500 | 509 | 503 |
| Pitt (1935) | 5 protanopes<br>6 deuteranopes | 494 | 497 | 496 | 495 | 507 | 500 |
| Hecht-Shlaer (1936) | 10 protanopes<br>12 deuteranopes | 492 | 508 | 498 | 491 | 525 | 510 |

end is expected. Wright[59] found that, among the 35 normal observers studied, variations in pigmentation made the point representing 5000°K vary from ($x=0.29$, $y=0.27$) for no ocular pigmentation to ($x=0.39$, $y=0.43$) for the most heavily pigmented observer studied. Table II shows the wave-lengths of the neutral points to be expected from protanopes and deuteranopes having these extremes of pigmentation, and it also shows for comparison the average and extreme wave-lengths of neutral points found by König,[35] Pitt,[52] and Hecht and Shlaer.[15] The König wave-lengths have been increased by 3 mμ for protanopes and 4 mμ for deuteranopes to take account of the chromaticity difference between the average daylight used by him and color temperature 5000°K. Pitt used Illuminant *B* for comparison in determining neutral points; the corrections for the chromaticity difference between this source and one at 5000°K are negligible and have not been applied. It will be noted from Table II that the König and Pitt data not only agree in average values with the neutral points read from Fig. 1, but also fall generally within the limits corresponding to those expected from the pigmentation range of Wright's 35 normal observers. The single deuteranope studied by König who falls slightly outside the pigmentation limits of Wright's 35 normal observers probably corresponds to nothing more than the error of random sampling.

One of the Hecht-Shlaer protanopes and 7 of their deuteranopes, however, have neutral points considerably higher than the maximum values corresponding to the most heavily pigmented of Wright's 35 normal observers. That 7 out of 12 deuteranopes are found to have heavier ocular pigmentation than any of 35 normal observers is too much to explain by the error of random sampling. This fact suggests strongly that the populations from which the observers were drawn are significantly different in ocular pigmentation. By this view we must be prepared to accept rather wide individual variations in pigmentation. For example, the deuteranope, verified by Hecht and Shlaer as having a neutral point at 525 mμ, must have an ocular pigmentation 8 or 9 times as heavy as the average found by Ludvigh and McCarthy[42] for ocular pigmentation of 62-year-old eyes, excluding the macular pigment. Table II also suggests the possibility that deuteranopes tend to have heavier pigmentation than protanopes or normals. It is concluded that individual variations in neutral point are ascribable to variations in ocular pigmentation (macula, lens, humors).

## IV. DERIVATION OF RESPONSE FUNCTIONS

It is worth noticing that the point ($x_p=0.747$, $y_p=0.253$) found from Pitt's data as the common crossing point of the protanopic chromaticity-confusion lines is very close to the long wave extreme of the spectrum ($x=0.735$, $y=0.265$) used as a primary in the OSA coordinate system.[57] Reference to Fig. 1a shows immediately that in spite of this close approach, the long wave extreme of the spectrum is not eligible to represent ($x_p$, $y_p$). For example, the line connecting the extremes of the spectrum cuts all of the protanopic chromaticity-confusion lines (dotted); that is, the protanopic chromaticity of the long wave extreme differs maximally from

that of the short wave extreme. But if the long wave extreme of the spectrum locus were itself taken as the point $(x_p, y_p)$, one interpretation would be that the two extremes of the spectrum have the same protanopic chromaticity. On this account the solid lines of Fig. 1a drawn in for comparison have been made copunctal at $(x_p=0.747,\ y_p=0.253)$. Although the choice of the long wave extreme of the spectrum to represent $(x_p, y_p)$ flatly contradicts the facts, it is of interest to inquire what response functions are generated by this choice because it bears on a question raised in 1798 by Dalton[3] and argued many times since, namely: Are protanopes red-blind?

Similarly it is worth noting that the point $(x_d=1.08,\ y_d=-0.08)$ estimated from Pitt's data as the most probable common crossing point of the deuteranopic chromaticity-confusion lines is fairly close to the $X$ primary $(x=1.00,\ y=0.00)$ of the standard I.C.I. coordinate system. The solid lines of Fig. 1b have been drawn through this point, and it is seen that with one exception the solid lines are either closely parallel to the nearest dotted line or cross a single dotted line. Since the separation of the dotted lines indicates the smallest chromaticity difference perceptible to Pitt's observers, it may be concluded that to set $x_d=1$ and $y_d=0$ is not inconsistent with these data; and, indeed, from a set of lines drawn through the point $(x_d=1.08,\ y_d=-0.08)$ it may be concluded that the former choice is scarcely less apt than the latter which was chosen as the best estimate possible. It will be of interest, therefore, to study both choices.

It is convenient to start this study by investigating the implications of setting $x_p+y_p=x_d+y_d=1$, or, stated another way, since $z=1-x-y$, the implications of setting $z_p=z_d=0$. All four points under consideration meet this requirement.

If $z_p=z_d=0$, from Eqs. (2a) and (2b), we find $k_7=k_8=0$, provided $x_p/y_p\neq x_d/y_d$, that is, provided $(x_p, y_p)$ and $(x_d, y_d)$ are different points; and from Eq. (3), we find that $k_9=1$. From Eq. (1) we may therefore write: $K=Z$. This conclusion is important. It says that since both the protanopic chromaticity-confusion lines and the deuteranopic-confusion lines are copunctal at points lying on the tangent to the spectrum locus at the long wave extreme, the $Z$ function of the standard observer is adequate to represent the $K$ function used in describing the chromaticity confusions of both deuteranope and protanope. König[37] from somewhat incomplete data[6] drew this conclusion, and v. Kries* deduced it from the identity of the $K$ function of protanope and deuteranope on the assumption that both are reduction forms of normal vision.

A further conclusion is possible from setting $z_p=z_d=0$. From (2b), (3), and (4), we may write three simultaneous equations with only three unknowns, thus:

$$k_1x_d+k_2y_d=0,\quad k_1+k_2+k_3=1,$$
$$k_3=-0.22k_1,$$

from which the coefficients, $k_1$ to $k_3$, defining the $W_d$ function by Eq. (1), are found to be:

$$k_1=-y_d/(x_d-0.78y_d),$$
$$k_2=x_d/(x_d-0.78y_d), \qquad (8a)$$
$$k_3=0.22y_d/(x_d-0.78y_d).$$

Thus it is seen that the $W_d$ function (warm curve of deuteranopes) is determined wholly by the coordinates of the point $(x_d, y_d)$ provided $z_d=z_p=0$, as they do for the four choices of interest.

Similarly, from Eqs. (2a), (3), and (4), we find that:

$$k_4=-y_p/(x_p-0.78y_p),$$
$$k_5=x_p/(x_p-0.78y_p), \qquad (8b)$$
$$k_6=0.22y_p/(x_p-0.78y_p).$$

Thus it is seen that the $W_p$ function (warm curve of protanopes) is determined wholly by the coordinates of the point $(x_p, y_p)$ again provided $z_d=z_p=0$.

The next step is to find the response functions resulting from setting $x_d=1$, $y_d=z_d=0$, as is justified from available data on chromaticity confusions of deuteranopes. From Eq. (8a), we find immediately that: $k_1=k_3=0$, and $k_2=1$. Hence from Eq. (1), $W_d=Y$. This conclusion is important. It says that since the deuteranopic chromaticity-confusion lines may be taken as copunctal at $x_d=1$, $y_d=z_d=0$, then the $Y$ function of the standard observer is adequate to

* Reference 39, p. 164.

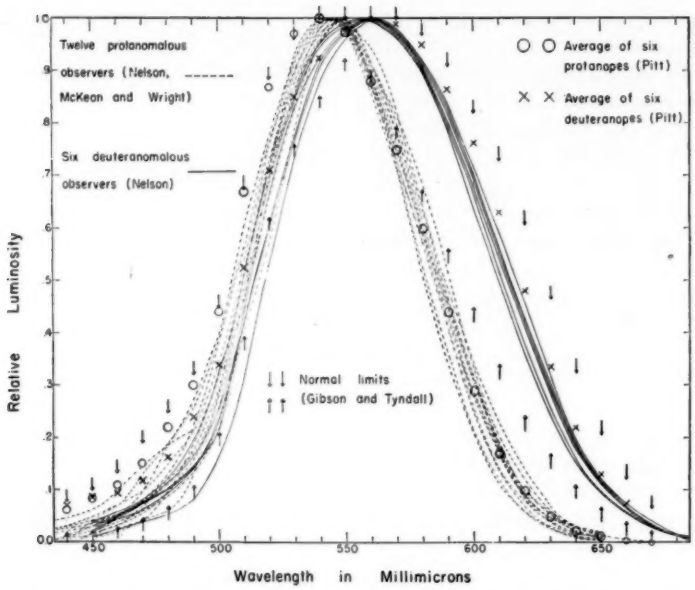



FIG. 2. Relative luminosity functions of anomalous trichromats (6 deuteranomalous, 12 protanomalous) compared to averages for dichromats and to limits for normal trichromats. Deuteranopic and deuteranomalous luminosity functions fall generally within normal limits; but protanopic and protanomalous luminosity functions fall outside of normal limits.

represent the $W_d$ function (warm curve of the deuteranope) for describing the chromaticity confusions of the deuteranope. The great convenience of being able to use two ($Z$ and $Y$) of the three functions representing the standard observer directly in deuteranopic chromaticity specifications should be emphasized. It means that the chromaticity of any color stimulus already specified in terms of the standard observer may be specified for an average deuteranope merely by leaving the $X$ specification out of account. That is, any two stimuli for which the $Y$ specification and the $Z$ specification bear the same ratio will be chromaticity matches for an average deuteranope.

The $W_p$ function is found from the values of $x_p=0.747$, $y_p=0.253$, through Eqs. (8b) and (1) to be:

$$W_p=-0.460X+1.359Y+0.101Z. \quad (9)$$

The $W_p$ function generated from the theoretically interesting but factually inadmissible values of $x_p=0.735$, $Y_p=0.265$ is found in a similar way to be:

$$W_p'=-0.503X+1.392Y+0.111Z. \quad (9a)$$

The $W_d$ function agreeing as closely to the Pitt average data as is possible ($x_d=1.08$, $y_d=-0.08$, $z_d=0.00$) is found from Eqs. (8a) and (1) to be:

$$W_d'=0.079X+1.061Y-0.017Z. \quad (10)$$

## V. LUMINOSITY FUNCTIONS OF RED-GREEN-BLIND OBSERVERS COMPARED TO THOSE OF NORMAL AND ANOMALOUS TRICHROMATS

We may now inquire into the relationship between the functions, $W_d$, $W_p$, $W_d'$, and $W_p'$, and the luminosity functions of dichromatic and trichromatic observers. Figure 2 shows by dotted lines the luminosity functions of eleven protanomalous observers studied by McKeon and Wright[45] and one studied by Nelson.[49] It also shows by solid lines the luminosity functions of six deuteranomalous observers studied by Nelson.[49] Although at the short wave end of the spectrum these groups of functions overlap and are not significantly different, the two groups divide quite sharply at the long wave end. In this respect they are like the groups of functions representing the two corresponding types of

dichromatic observer, the deuteranope and the protanope. The circles represent the average of six protanopic luminosity functions evaluated by Pitt,[52] and the crosses represent his average of six deuteranopic luminosity functions. The arrows indicate the maximum and minimum values of relative luminosity found by Gibson and Tyndall[11] among their 37 completely studied normal observers. Although because of differences in experimental conditions no highly precise intercomparison is possible between the data shown for anomalous and dichromatic observers and these normal limits, nevertheless Fig. 2 indicates that the deuteranopic and deuteranomalous observers alike possess luminosity functions generally well within normal limits; but both protanomalous observers and protanopic observers possess luminosity functions that are abnormally low at the long wave end.

Figure 2 tends to indicate also that protanomalous observers possess abnormally narrow luminosity functions even compared to protanopes, and that deuteranopes possess luminosity functions abnormally high between 570 and 610 mμ with deuteranomalous observers intermediate in this respect; but because of the small number of abnormal observers tested, these indications may fail to be significant. The experimental conditions (field size, field luminance, surrounding-field luminance) and techniques (adaptation, control of fixation, length of observing time) govern the participation of the rod mechanism in determining these functions. In any one study the luminosity curves may be thrown more or less to the short wave side because of partial participation of the rods, either intended or inadvertent. The intertwining of the curves at the short wave end is probably to be ascribed chiefly to individual variation in the amount of yellow or brown pigmentation of the eye media, such as degree of macular pigmentation.

Since from Fig. 2 both deuteranopic and deuteranomalous luminosity functions fall generally within the limits of 37 normal observers, it is just as valid to predict the brightness judgments of these abnormal observers from the standard luminosity function as it is those of an

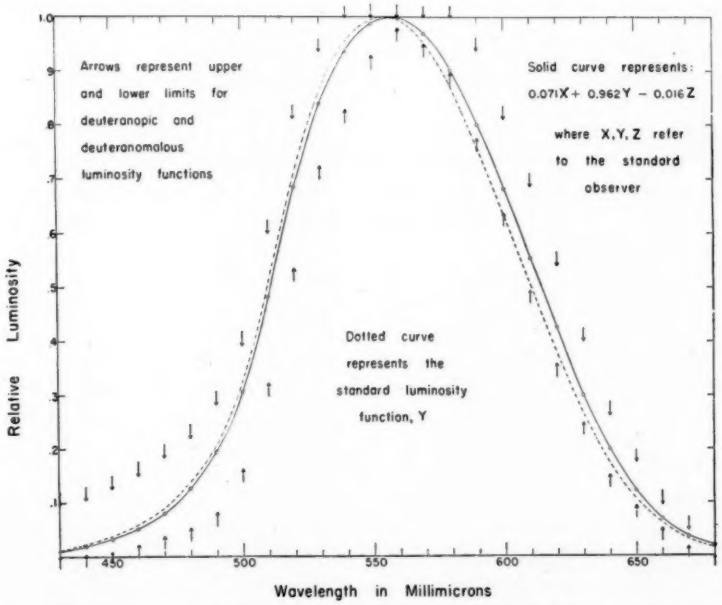

Arrows represent upper and lower limits for deuteranopic and deuteranomalous luminosity functions
Solid curve represents: 0.071X + 0.962Y − 0.016Z where X, Y, Z refer to the standard observer
Dotted curve represents the standard luminosity function, Y
Relative Luminosity
Wavelength in Millimicrons
450
500
550
600
650

FIG. 3. Comparison of the functions $W_d$ (standard luminosity function $Y$) and $W_d'$ (solid curve) with upper and lower limits for deuteranopic and deuteranomalous luminosity functions (arrows). The standard luminosity function is seen to fall satisfactorily within these limits except for the wave-length region between 570 and 610 mμ. The function $W_d'$ yields slightly better agreement.

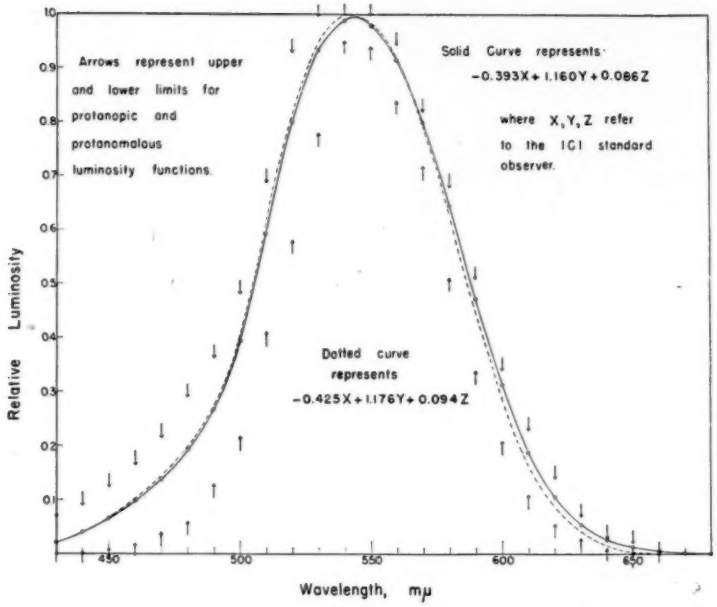



FIG. 4. Comparison of the functions $W_p$ (solid curve) and $W_p'$ (dotted curve) with upper and lower limits for protanopic and protanomalous luminosity functions (arrows). Both functions are seen to fall satisfactorily within these limits except that $W_p'$ is perhaps too low for wave-lengths greater than 650 mμ.

observer chosen at random from the large population group to be classed as of normal color vision. Figure 3 compares the standard luminosity function $Y=W_d$ (dotted line) with the upper and lower limits (arrows) for the deuteranopic and deuteranomalous luminosity functions referred to in Fig. 2. The standard luminosity function is seen to fall satisfactorily within these limits except for the wave-length region between 570 and 610 mμ. Studies by Sloan[56] corroborated by the recent work of Walters and Wright[58] show that the shape of the luminosity function within this wave-length region is particularly subject to variation with experimental conditions. It is therefore perhaps reasonable to ascribe the whole of this small discrepancy to variation in experimental conditions.

Figure 3 also shows the function $W_d'$ [Eq. (10)] adjusted to unit maximum as a solid line. This function represents available data on deuteranopic luminosity slightly better than the standard luminosity function $Y=W_d$, just as, combined with the $K$ function, it represents deuteranopic chromaticity data slightly better. There is therefore some basis for arguing, as in the most common form of Young-Helmholtz theory, that the normal luminosity function should be thought of as consisting of three components, a red, a green, and a violet.* Whether this possible justification of the usual Young-Helmholtz view would be borne out by a study of an adequate sample of normal and deuteranopic observers remains to be decided by further extensive experiment. The two functions, $W_d$ and $W_d'$, differ so little, however, that it would rarely be of practical importance to use one rather than the other. Because of its greater simplicity it seems better to take $W_d=Y$ as the function representing deuteranopic and deuteranomalous luminosity.

Figure 4 compares the functions $0.854W_p$ and $0.845W_p'$ with the upper and lower limits (arrows) for the protanopic and protanomalous luminosity functions referred to in Fig. 2. The constants 0.854 and 0.845 serve to adjust the functions approximately to unit maximum, and

* An account of dichromatic vision following this view has recently been worked out by Pitt (reference 53).

from Eqs. (9) and (9a), respectively, yield:

$$0.854W_p = -0.393X + 1.160Y + 0.086Z$$
(solid curve),

$$0.845W_p' = -0.425X + 1.176Y + 0.094Z$$
(dotted curve).

It is seen that both functions fall satisfactorily between the limits at least for wave-lengths shorter than 650 mμ. The function $W_p$, therefore, not only represents adequately, when combined with the $K$ function, the chromaticity confusions of the average protanope, but also, taken by itself, his brightness judgments. The function $W_p'$, although it grossly fails, when combined with the $K$ function, to represent the chromaticity confusions of the protanope, nevertheless succeeds in representing adequately his brightness judgments except for wave-lengths greater than 650 mμ. This rather minor difference in luminosity function is all that differentiates red-blind vision (luminosity zero for wave-lengths greater than 700 mμ) from the experimentally determined facts of protanopia. The old name, red blindness, is therefore nearly but not quite accurate.

By way of contrast it may be pointed out that the corresponding old name, green blindness, for the type of defect known since v. Kries[46] as deuteranopia is almost wholly wrong. Since the standard luminosity function represents the brightness judgments of deuteranopic observers just as closely as those of a considerable fraction of normal observers, and is far from being zero anywhere within the spectrum region which ordinarily appears green (490 to 550 mμ), deuteranopes are far from being insensitive to the green part of the spectrum. It is true that protanopia and deuteranopia can logically be described as red blindness and green blindness, respectively, in terms of the Young-Helmholtz theory, but in a purely descriptive nontheoretical sense, no observer has ever been found to be quite red-blind, and none even remotely deserving the name, green-blind.

The following functions therefore may be taken to be representative of normal, deuteranopic and protanopic vision:

$$\begin{aligned} K &= Z, \\ W_p &= -0.460X + 1.359Y + 0.101Z, \\ W_d &= Y, \end{aligned} \qquad (11)$$

$K$ and $W_d$ for deuteranopic vision, $K$ and $W_p$ for protanopic vision, and all three for normal vision. It is also true, although no adjustments other than Eq. (4) have been made for it nor any detailed quantitative comparisons presented, that $W_p$ and $W_d$ taken together closely represent tritanopic vision.[31,33,38,47] Table III gives these three functions for 10-mμ intervals of the visible spectrum. The function $W_p$ is taken as the luminosity function of a typical protanope, the function $W_d$ is taken as the luminosity function of a typical deuteranope as well as the

TABLE III. Response functions for protanopic ($W_p$, $K$) and deuteranopic ($W_d$, $K$) vision.

| Wave-length in mμ | $W_d$ | $W_p$ | $K$ |
|---|---|---|---|
| 380 | 0.00004 | 0.00006 | 0.0065 |
| 390 | .00012 | .00024 | .0201 |
| 400 | .0004 | .00082 | .0679 |
| 410 | .0012 | .00257 | .2074 |
| 420 | .0040 | .00883 | .6456 |
| 430 | .0116 | .0251 | 1.3856 |
| 440 | .0230 | .0476 | 1.7471 |
| 450 | .0380 | .0759 | 1.7721 |
| 460 | .0600 | .1163 | 1.6692 |
| 470 | .0910 | .1638 | 1.2876 |
| 480 | .1390 | .2270 | .8130 |
| 490 | .2080 | .3150 | .4652 |
| 500 | .3230 | .4642 | .2720 |
| 510 | .5030 | .6953 | .1582 |
| 520 | .7100 | .9437 | .0782 |
| 530 | .8620 | 1.0997 | .0422 |
| 540 | .9540 | 1.1650 | .0203 |
| 550 | .9950 | 1.1537 | .0087 |
| 560 | .9950 | 1.0791 | .0039 |
| 570 | .9520 | .9434 | .0021 |
| 580 | .8700 | .7610 | .0017 |
| 590 | .7570 | .5568 | .0011 |
| 600 | .6310 | .3690 | .0008 |
| 610 | .5030 | .2224 | .0003 |
| 620 | .3810 | .1248 | .0002 |
| 630 | .2650 | .0646 | .0000 |
| 640 | .1750 | .0318 | .0000 |
| 650 | .1070 | .0150 | .0000 |
| 660 | .0610 | .0070 | .0000 |
| 670 | .0320 | .0033 | .0000 |
| 680 | .0170 | .00157 | .0000 |
| 690 | .0082 | .00070 | .0000 |
| 700 | .0041 | .00035 | .0000 |
| 710 | .0021 | .00018 | .0000 |
| 720 | .00105 | .000089 | .0000 |
| 730 | .00052 | .000044 | .0000 |
| 740 | .00025 | .000021 | .0000 |
| 750 | .00012 | .000010 | .0000 |
| 760 | .00006 | .000005 | .0000 |
| 770 | .00003 | .000003 | .0000 |

normal observer (and perhaps also that of a typical tritanope),[7,8,34] and the function $K$ is unassociated with luminosity.

The representation of normal vision by $K$, $W_d$, and $W_p$ is just as adequate as that afforded by the functions $Z$, $Y$, and $X$ of the standard I.C.I. coordinate system, and indeed functions closely like these were proposed by the author in 1931 for international adoption through the American representative, Mr. Irwin G. Priest, but were rejected because the corresponding chromaticity diagram gave considerably less uniform chromaticity scales than the coordinate system ($X$, $Y$, $Z$) finally adopted. There is therefore nothing to be gained by supplanting the present $X$ function by $W_p$ to represent normal vision, and this substitution is not recommended.

The representation of deuteranopic vision by $K$ and $W_p$, and protanopic vision by $K$ and $W_p$, is valid in the same way as representation of normal vision by $X$, $Y$, and $Z$. No single normal observer, or protanope, or deuteranope, can be found except by accident who conforms to the standard representation of the respective group, but the departures of each single individual from the standard may be expected to be smaller than some of the individual differences within the group.

TABLE IV. Chromaticity coordinates for deuteranopic and protanopic vision.

| Wave-length in mμ | Chromaticity coordinates: Protanopia $w_p$ | Protanopia $k_p$ | Deuteranopia $w_d$ | Deuteranopia $k_d$ |
|---|---|---|---|---|
| 380 | 0.012 | 0.988 | 0.006 | 0.994 |
| 430 | .018 | .982 | .008 | .992 |
| 450 | .041 | .959 | .021 | .979 |
| 460 | .065 | .935 | .035 | .965 |
| 470 | .113 | .887 | .066 | .934 |
| 475 | .156 | .844 | .098 | .902 |
| 480 | .218 | .782 | .146 | .854 |
| 485 | .301 | .699 | .216 | .784 |
| 490 | .404 | .596 | .309 | .691 |
| 495 | .518 | .482 | .423 | .577 |
| 500 | .631 | .369 | .543 | .457 |
| 505 | .730 | .270 | .657 | .343 |
| 510 | .815 | .185 | .761 | .239 |
| 515 | .881 | .119 | .845 | .155 |
| 520 | .923 | .077 | .901 | .099 |
| 530 | .963 | .037 | .953 | .047 |
| 540 | .982 | .018 | .979 | .021 |
| 550 | .992 | .007 | .991 | .009 |
| 600 | .998 | .002 | .999 | .001 |
| 650 | 1.000 | .000 | 1.000 | .000 |
| 750 | 1.000 | .000 | 1.000 | .000 |

## VI. USE OF THE RESPONSE FUNCTIONS

The response functions $K$, $W_p$, and $W_d$ have been used by the ISCC committee on color-blindness* to find whether or not two colors will be confused by an average protanope or an average deuteranope. Such use will be illustrated by solution of three problems; first, what is the influence of the choice of standard neutral source on the dichromatic neutral points in the spectrum, second, how much darker to an average protanope are normally reddish colors, and third, what arrangements of the colors of the Farnsworth dichotomous test B-20[9] are characteristic of average deuteranopic and protanopic vision.

### A. Wave-Length of Neutral Point

The determination of the neutral point in the spectrum of a dichromatic observer involves setting up a chromaticity match between the standard neutral source (I.C.I. Illuminant $A$, I.C.I. Illuminant $C$, blackbody at 5000°K, and so forth) and some part of the spectrum; see Table II. Deuteranopic and protanopic chromaticity is conveniently specified by chromaticity coordinates $w_d$, $k_d$ and $w_p$, $k_p$, the first pair for deuteranopic vision, the second for protanopic. The definitions of these coordinates and their equivalents in terms of $X$, $Y$, $Z$, from Eq. (11) are as follows:

$$w_d = W_d/(W_d+K) = Y/(Y+Z), \qquad k_d = K/(W_d+K) = Z/(Y+Z); \tag{12a}$$

$$w_p = W_p/(W_p+K) = \frac{-0.460X+1.359Y+0.101Z}{-0.460X+1.359Y+1.101Z}, \tag{12b}$$

$$k_p = K/(W_p+K) = Z/(-0.460X+1.359Y+1.101Z).$$

* Committee on color-blindness tests, Inter-Society Color Council, Co-Chairmen: Dr. LeGrand H. Hardy, and the author.

TABLE V. Protanopic and deuteranopic neutral points in the spectrum derived from the dichromatic chromaticity coordinates $w_p$ and $w_d$.

| Standard source | I.C.I. chromaticity coordinates $x$ | $y$ | Dichromatic chromaticity coordinates $w_p$ | $w_d$ | Wave-length of the neutral point, mμ $\lambda_p$ | $\lambda_d$ |
|---|---|---|---|---|---|---|
| I.C.I. Illuminant *A* | 0.4476 | 0.4075 | 0.7144 | 0.7377 | 504.0 | 508.9 |
| I.C.I. Illuminant *B* | .3484 | .3516 | .5369 | .5396 | 495.8 | 499.8 |
| Complete radiator at 5000°K | .3445 | .3512 | .5346 | .5358 | 495.7 | 499.7 |
| I.C.I. Illuminant *C* | .3101 | .3163 | .4652 | .4585 | 492.7 | 496.4 |
| Illuminant *D** | .2999 | .3120 | .4559 | .4457 | 492.3 | 495.8 |
| Illuminant *S*† | .2319 | .2318 | .3286 | .3018 | 486.4 | 489.6 |

* Illuminant D is produced by an incandescent lamp and Corning Daylite glass filter (reference 32, p. 60), and has a nearest color temperature (reference 28) of about 7500°K.

† Illuminant S was found (reference 32, p. 58) by weighting Abbot's "sun-outside-atmosphere" energy data by the inverse $\lambda^4$ scattering relation, and has been designated as "limit blue sky."

Table IV shows these coordinates for the various parts of the spectrum, particularly the portions near 495 mμ where chromaticity for these types of dichromatic vision varies most rapidly with wave-length. Since from their definitions the two coordinates sum to unity, only one of them is required to specify the dichromatic chromaticity of a stimulus.

In order to find the neutral point in the spectrum of, say, an average protanope for a standard neutral source specified in the standard I.C.I. system by $X_0$, $Y_0$, $Z_0$, it is necessary only to compute $w_p$ from Eq. (12b) and find by interpolation among the values of $w_p$ for the spectrum given in Table IV the wave-length for the corresponding value. This wave-length corresponds to the same chromaticity as the standard neutral source and is therefore the desired neutral point. This method is the analytic equivalent of a graphical method using the chromaticity diagram of the standard system; see Fig. 1. The graphical method is to draw a straight line from the copunctal point through the point representing the chromaticity of the standard neutral source and extend it until it cuts the spectrum locus. The wave-length corresponding to this intersection is the desired neutral point in the dichromatic spectrum.

Table V gives the standard chromaticity specifications, $x$ and $y$, for several sources that might be taken as standard neutrals, the values of $w_p$ and $w_d$ for these sources, and the corresponding neutral points found by the analytical method from Table IV. It will be noted that the change in neutral point depending upon choice of standard neutral source is considerably greater than the difference between the neutral points of the typical protanope and typical deuteranope for any one source.

### B. Protanopic Luminous Reflectance

The second problem is to find how much darker to an average protanope are normally reddish colors. For spectrum colors the solution is given simply by comparing $W_p$ with $Y=W_d$ in Table III. It is seen, for example, that relative to an equal-energy source, an average protanope finds the spectrum from 700 to 770 mμ to have less than 10 percent of the normal luminosity. It could be said therefore in a sense that he is more than 90 percent blind to the long wave end of the spectrum; but we would also have to say in the same sense that he is supersensitive by a factor of 2 for the spectral region near 450 mμ. It is more convenient to say that the luminosity function of the average protanope is shifted about 10 mμ toward the short wave end of the spectrum relative to normal.

If the color is that of a surface it is customary to estimate its lightness or darkness from the luminous reflectance of the surface for I.C.I. Illuminant *C* relative to magnesium oxide. For magnesium oxide, $X=0.9804$, $Y=1.0000$, and $Z=1.1812$, and from Eq. (9), $W_p=1.0274$. Protanopic luminous reflectance relative to magnesium oxide is therefore equal to $W_p/1.0274$ or $0.9733W_p$. It is convenient to represent on the $(x, y)$ chromaticity diagram the dependence of protanopic luminous reflectance of a surface on the normal chromaticity of its color.

If $Q$ be the ratio of the protanopic luminous reflectance of a surface to its normal luminous

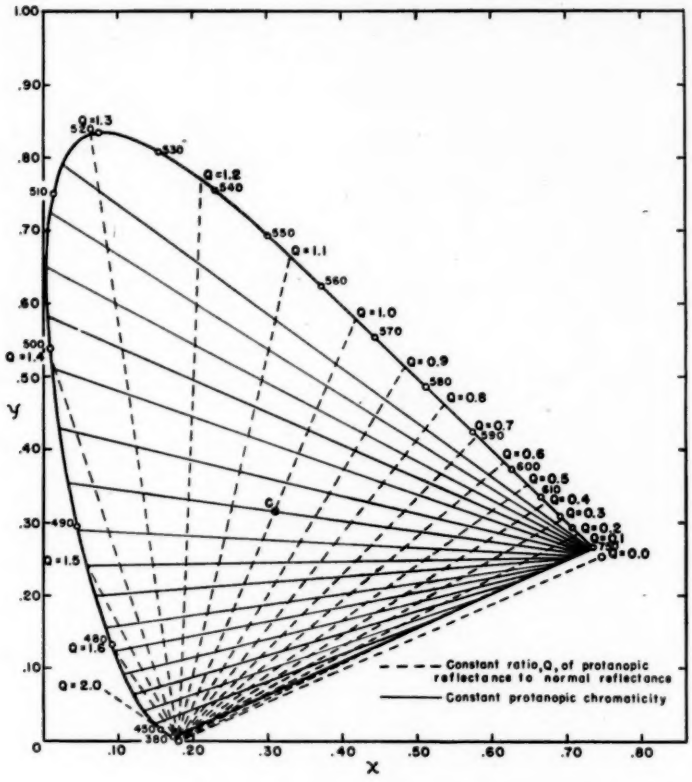



FIG. 5. Relation of protanopic to normal vision. The solid lines (as in Fig. 1) represent constant protanopic chromaticity; the dotted lines represent constant values of the ratio $Q$ of protanopic luminous reflectance to normal luminous reflectance. Both reflectances are taken relative to that of magnesium oxide and refer to I.C.I. standard Illuminant $C$. Note that the line (dotted) for $Q=1$ passes through the point, $x=0.310$, $y=0.316$, representing Illuminant $C$. Note also that the line for $Q=0$ passes through the copunctal point, $x=0.747$, $y=0.253$; this line could be called the protanopic alychne (lightless line), just as the line for $Q=\infty$ [$y=0$, see Eq. (13)] has sometimes been called the normal alychne.

reflectance $Y$, we may write:

$$Q=0.9733W_p/Y$$
$$=0.9733(-0.460X+1.359Y+0.101Z)/Y$$
$$=0.9733(-0.460x+1.359y+0.101z)/y,$$

whence it is found:

$$x=y(Q-1.224)/(-0.546)+0.180. \quad (13)$$

Examination of Eq. (13) shows that it represents a family of straight lines on the $(x, y)$ chromaticity diagram passing through the point ($x=0.180$, $y=0.000$). For $Q=1$, the line passes through the point representing Illuminant $C$ ($x=0.3101$, $y=0.3163$). All surfaces whose chromaticities under Illuminant $C$ are represented on this line have identical protanopic and normal luminous reflectances. For $Q=2.0$, the line passes close to 450 m$\mu$; all surfaces represented on this line have protanopic luminous reflectance twice the normal luminous reflectance. For $Q=0.10$, the line passes close to 650 m$\mu$; all surfaces represented on this line have protanopic luminous reflectance equal to 10 percent of the normal luminous reflectance. Figure 5 shows a number of these lines of constant $Q$ (dotted) on the $(x, y)$ chromaticity diagram; it indicates completely and quantitatively how much darker to an average protanope are normally reddish colors. It also

shows the lines of protanopic chromaticity confusion; it therefore gives a convenient, concise and quantitative summary of the connection between average protanopic vision and the I.C.I. standard observer.

Two of the series of straight lines of constant $Q$ have special interest. One of them ($Q=\infty$) corresponds to $y=0$, sometimes called the alychne[25, 55] because on it are represented imaginary chromaticities which are unassociated with luminosity. The only way to make the quotient $Q$ become indefinitely large is to make the denominator $Y$ equal zero; that is, by choosing a chromaticity represented on the alychne. The other line of special interest is found by setting $Q=0$, which defines the line:

$$x=y(1.224/0.546)+0.180. \qquad (14)$$

This line could well be called the protanopic alychne. Note that it passes through the copunctal point ($x=0.747$, $y=0.253$) of the lines representing the chromaticities confused by the average protanope, and that it does not cross any portion of the chromaticity diagram representing real surfaces.

The copunctal point for the lines of constant $Q$ is found from Eq. (11) also to be the point for which $W_d=W_p=0$. If, as seems at least approximately correct, $W_d$ and $W_p$ taken together represent average tritanopic vision, then it can be said that the lines of constant ratio between protanopic and normal reflectance are also chromaticity-confusion lines for tritanopic vision.

### C. Farnsworth Dichotomous Test (B-20)

The third problem is to find the characteristic deuteranopic and protanopic arrangement of the colors of the Farnsworth dichotomous test B-20. Table VI gives the Munsell hue of the colors used, Munsell value and chroma being constant at 5/2, their tristimulus specifications ($X$, $Y$, $Z$) for I.C.I. Illuminant $C$, one of the distimulus specifications $W_p$ multiplied by the constant $0.973_3$ to make it correspond to protanopic reflectance relative to MgO, and the dichromatic chromaticity coordinates $w_p$ and $w_d$. The fraction $W_d/(W_p+W_d)$, identified by the notation $w_t$, is also given in order to test whether this fraction correlates with the arrangement of the colors found by Farnsworth to be characteristic of a tritanope. It is first to be noted from the relative constancy of the daylight reflectance, $W_d=Y$, and the protanopic reflectance, $0.9733W_p$, for these colors that the papers used in the Farnsworth dichotomous test have nearly the same reflectances for Illuminant $C$ not only for the standard observer and typical deuteranope

TABLE VI. Tristimulus and distimulus specifications of the colors of the Farnsworth dichotomous test for color vision.

| Serial No. | Munsell hue (at 5/2) | Tristimulus specifications $X$ | $Y$ | $Z$ | Protanopic reflectance $0.9733\,W_p$ | Distimulus chromaticity coordinates $w_p$ | $w_d$ | $w_t$ |
|---|---|---|---|---|---|---|---|---|
| 1 | 10*B* | 0.2121 | 0.2229 | 0.3292 | 0.2321 | 0.4201 | 0.4037 | 0.4831 |
| 2 | 5*B* | .1749 | .1917 | .2831 | .2030 | .4242 | .4037 | .4789 |
| 3 | 10*BG* | .1820 | .2043 | .2790 | .2162 | .4432 | .4227 | .4791 |
| 4 | 5*BG* | .1782 | .2077 | .2538 | .2199 | .4709 | .4501 | .4790 |
| 5 | 5*G* | .1618 | .1888 | .2017 | .1972 | .5011 | .4835 | .4824 |
| 6 | 10*GY* | .1766 | .2017 | .1873 | .2061 | .5307 | .5185 | .4878 |
| 7 | 5*GY* | .1908 | .2097 | .1804 | .2096 | .5442 | .5376 | .4933 |
| 8 | 10*Y* | .2135 | .2292 | .1798 | .2253 | .5628 | .5604 | .4975 |
| 9 | 5*Y* | .2057 | .2117 | .1795 | .2056 | .5406 | .5412 | .5006 |
| 10 | 10*YR* | .2170 | .2143 | .1780 | .2038 | .5405 | .5463 | .5058 |
| 11 | 5*YR* | .2029 | .1927 | .1755 | .1813 | .5149 | .5234 | .5084 |
| 12 | 10*R* | .2364 | .2185 | .2055 | .2034 | .5042 | .5153 | .5111 |
| 13 | 5*R* | .2158 | .1954 | .2041 | .1818 | .4779 | .4891 | .5113 |
| 14 | 10*RP* | .2426 | .2229 | .2453 | .2103 | .4684 | .4761 | .5077 |
| 15 | 5*RP* | .2304 | .2138 | .2559 | .2048 | .4512 | .4552 | .5040 |
| 16 | 10*P* | .2256 | .2092 | .2751 | .2027 | .4309 | .4320 | .5011 |
| 17 | 5*P* | .2070 | .1959 | .2818 | .1942 | .4145 | .4101 | .4954 |
| 18 | 10*PB* | .2199 | .2118 | .3301 | .2140 | .3998 | .3908 | .4906 |
| 19 | 7.5*PB* | .2055 | .2035 | .3036 | .2071 | .4121 | .4013 | .4888 |
| 20 | 5*PB* | .2052 | .2087 | .3277 | .2164 | .4042 | .3891 | .4842 |

($W_d = Y$), but also for the typical protanope 0.9733 ($W_p$), the variation being between 0.2321 and 0.1818 which corresponds to only 0.55 of a Munsell value step.[50] Since these papers are exhibited with a dark gray border which interferes with perception of lightness differences more than it does chromaticity differences,* the basis for arranging the colors in order must be chiefly according to chromaticity. The characteristic protanopic arrangement is therefore to be found in accord with $w_p$, and the characteristic deuteranopic arrangement according to $w_d$.

Table VII shows the serial order according to $w_p$ compared to that for a protanope reported by Farnsworth. That is, color number 8 was picked by the Farnsworth protanope as the warmest of the 20 colors and it also has the highest value of $w_p$ (0.5628 from Table V); so number 8 heads both the second and third column of Table VI. A similar comparison is given between the arrangements according to $w_d$ and a deuteranope reported by Farnsworth, and between $w_t$ and the tritanope reported by Farnsworth. It is evident from Table VII that the correlations indicated by these comparisons are high; that is, the results chosen by Farnsworth to be typical of protanopia, deuteranopia, and tritanopia agree well though not perfectly with the chromaticity arrangements according to $w_p$, $w_d$, and $w_t$, respectively.

TABLE VII. Comparison of the serial orders of the colors of the Farnsworth dichotomous test according to $w_p$, $w_d$, and $w_t$ with the orders reported by Farnsworth for a protanope (red-bluegreen confuser), and deuteranope (green-redpurple confuser), and a tritanope (violet-greenishyellow confuser).

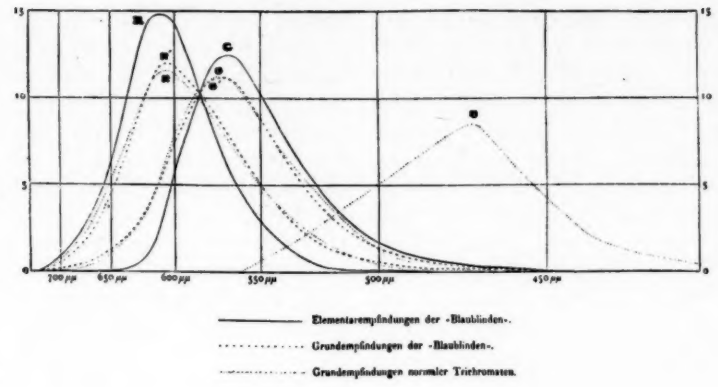
| Normal order (based on hue) | Serial numbers in order from warmest to coldest dichromatic color according to: | | | | | |
|---|---|---|---|---|---|---|
| | $w_p$ | Farnsworth protanope | $w_d$ | Farnsworth deuteranope | $w_t$ | Farnsworth tritanope |
| 1 | 8 | 8 | 8 | 10 | 13 | 13 |
| 2 | 7 | 10 | 10 | 9 | 12 | 14 |
| 3 | 9 | 7 | 9 | 8 | 11 | 15 |
| 4 | 10 | 6 | 7 | 7 | 14 | 12 |
| 5 | 6 | 9 | 11 | 11 | 10 | 11 |
| 6 | 11 | 11 | 6 | 12 | 15 | 16 |
| 7 | 12 | 5 | 12 | 6 | 16 | 10 |
| 8 | 5 | 12 | 13 | 13 | 9 | 9 |
| 9 | 13 | 13 | 5 | 14 | 8 | 17 |
| 10 | 4 | 4 | 14 | 5 | 17 | 18 |
| 11 | 14 | 3 | 15 | 15 | 7 | 8 |
| 12 | 15 | 15 | 4 | 16 | 18 | 19 |
| 13 | 3 | 14 | 16 | 4 | 19 | 7 |
| 14 | 16 | 2 | 3 | 17 | 6 | 20 |
| 15 | 2 | 16 | 17 | 3 | 20 | 1 |
| 16 | 1 | 17 | 2 | 2 | 1 | 6 |
| 17 | 17 | 1 | 1 | 18 | 5 | 2 |
| 18 | 19 | 19 | 19 | 20 | 3 | 5 |
| 19 | 20 | 20 | 18 | 19 | 4 | 4 |
| 20 | 18 | 18 | 20 | 1 | 2 | 3 |

Failure of the agreement to be perfect might be caused by failure of the tristimulus specifications reported for certain samples of the Munsell papers to apply exactly to the particular samples used in the Farnsworth test B-20, to careless arrangement of the colors by the dichromatic observer, or to arrangement on a basis other than chromaticity, or to significant difference between the visual system of the dichromat and that represented by the pairs of functions, $K$, $W_p$, and $W_d$ taken here to represent the average protanope and deuteranope. By plotting the orders given in Table VII according to $w_p$, $w_d$, and $w_t$ on the Farnsworth chart for analyzing results on the dichotomous test (see his Fig. 15), it is discovered that most of the minor differences are ascribable to a real difference between the typical Farnsworth dichromats and the corresponding hypothetical observers defined by the functions $K$, $W_p$, and $W_d$. The axes of confusion for the hypothetical observers fall, however, within the limits indicated by Farnsworth (see his Fig. 15) as embracing the three types of dichromat delineated by the dichotomous test.

Although this corroboration was expected for protanopic and deuteranopic vision because of the firm basis for choosing the functions $K$, $W_p$, and $W_d$, it is interesting to note that substantially the same agreement is found for tritanopic vision. The discrepancies are of the sort to be ascribed to individual differences. It is evident that the Farnsworth test is well adapted for surveys of the characteristics of dichromats; a statistical treatment of an extended body of data found by this test would serve to determine whether the functions proposed here for dichromatic vision describe, as was intended, visual systems intermediate to actual observers of the corresponding type; also whether the functions $W_p$ and $W_d$, taken together, describe a type of tritanopic vision intermediate to actual tritanopes.

* Reference 29, p. 425.

(a)




(b)

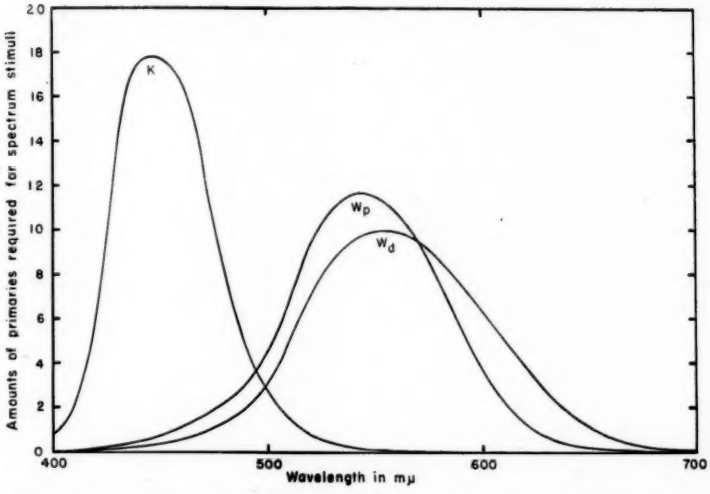



FIG. 6. Comparison of the functions $W_d$, $W_p$, and $K$ with the "fundamental-sensation" curves (dotted) derived by König in 1897. If allowance is made for the descending and expanding wave-length scale and for the energy distribution of gas light used by König, it will be noted that there is no essential difference between the two sets of functions.

## VII. THEORETICAL IMPLICATIONS OF THE RESPONSE FUNCTIONS

The functions $W_d$, $W_p$, and $K$ defined in Eq. (11) have theoretical implications the validity of which in no way interferes with the practical applications just discussed. It is convenient to bring out these implications by indicating the relation of these functions to the later theoretical views of Arthur König, to the recent dominator-modulator theory of Granit, and to the zone-theory of G. E. Müller.

### A. The König Theories

In 1886 and in 1893 Arthur König published accounts of his monumental work on the color systems of normal and abnormal observers.[36,37] The coordinate system chosen by him to represent his experimental results satisfies Eqs. (1) and (2) but does not satisfy Eq. (4) since one of the "fundamental sensations" was taken as unitary blue. In 1897, however, König published an important paper on tritanopia or "blue-blindness" in which choice of unitary blue as a

"fundamental sensation" was given up. He says in a footnote,* "I will immediately remark here that regarding the quality of this third fundamental sensation of the Young-Helmholtz color theory—whether it be blue or violet—I can still pass no definite judgment. In any case, however, I am now inclined to move it far closer to violet than was the case in the year 1886 (compare Sitz. Akad. Wiss. Berlin, July 29, 1886)."

As a result of this change in choice of third primary the "fundamental-sensation" curves bear a close resemblance to the functions $W_d$, $W_p$, and $K$. Figure 6a is a reproduction of König's Fig. 2 shown for comparison with Fig. 6b which is a plot of the functions $W_d$, $W_p$, and $K$. The dotted curves of Fig. 6a (the fundamental-sensation curves) are plotted against a descending and expanding wave-length scale, the curves are adjusted to equal area on this non-uniform scale, and refer to the dispersion spectrum of gaslight. Apart from these differences, however, it could be said that the functions $W_d$, $W_p$, and $K$ are modern evaluations of the functions taken in 1897 by König to represent the "fundamental-sensation" curves of the eye. From the standpoint of the König theory, therefore, it is but to be expected that $W_d$ and $W_p$ taken together describe the color vision of a tritanope. It is no small tribute to König that the simplest representation of normal and dichromatic vision possible at the present time so closely resembles the picture given by him more than forty years ago.

These later König views have so far received much less attention than his earlier evaluation of the "fundamental-sensation" curves which were included in the second edition of Helmholtz' *Physiological Optics*,[16] made the basis of a new color theory by Ladd-Franklin, and reinserted in that form in the English translation of the third German edition of Helmholtz.[18] Although he apparently did not draw attention to it in any publication, König immediately recognized that the distribution of the fundamental "red" sensation throughout the spectrum was closely the same as the luminosity distribution, and he discussed this consequence of his new view with, among others, his brilliant former pupil, Christine Ladd-Franklin, and J. v. Kries.* The then current view of the Young-Helmholtz theory was that luminosity should be thought of as made up of the sum of three components, a red, a green, and a blue or violet component. König's new view that all or nearly all of the luminosity might reside in a single one (red) of the components met with no favor at that time. This view as well as his previous views were omitted entirely from the third edition of Helmholtz[17] prepared in 1911 by Gullstrand, Nagel, and v. Kries. However, the completeness with which König in 1897 anticipated the present description of dichromatic and trichromatic vision is strikingly shown in the following footnote which appears in a note specially prepared in 1924 by v. Kries for insertion in the English edition of Helmholtz,** "König long ago called attention incidentally to the curious fact that the luminosity values of the colours . . . turn out to be very nearly the same function of the wave-length for deuteranopes and persons with normal vision. Since in the case of deuteranopes light of long wave-lengths has no effect except on the red component, we must suppose also that this action depends on the wave-length in the same way. But then the further result is that the distribution of luminosity is not affected, or at least almost inappreciably affected, by the addition of the green component by which the deuteranopic visual organ is converted into a normal organ; in other words, that in the case of normal vision also the luminosity goes practically hand in hand with the action of the red component."

### B. Dominator-Modulator Theory of Granit

In 1943 Granit[12] gave a brief outline of a physiological theory of color perception based upon his oscillographic studies of the responses of isolated fibers from the optic nerves of animals (cat, guinea pig, frog, snake, rat). In the light-adapted eyes of these animals the simple spectral sensitivity curves recorded with his micro-electrode technique were found to be of two types: (a) broad absorption bands, called *dominators*,

* Reference 38, p. 716; reference 38a, p. 396.

* Communicated privately to the author by Dr. Ladd-Franklin in 1928.

** Reference 18, p. 412.

and (b) narrow bands, called *modulators*. The dominator of the light-adapted animal eye was found to have its maximum in the spectral region around 560 mμ. The form and spectral location of the dominator were found to be practically identical with the average curve obtained from massed receptors in the light-adapted eyes of the same species. This and its good correspondence with respect to form and location with the luminosity curve of the light-adapted human eye led Granit to the conclusion that the dominator is responsible for the sensation of brightness, which thus is taken as our dominant impression. Modulation of the dominant impression of brightness to color would seem to be the task of the much rarer modulators which occupy narrow bands of sensitivity in three preferential regions around 580 to 600 mμ, 520 to 540 mμ, and 450 to 470 mμ.

It will be noted from Fig. 6b that the functions $W_d$, $W_p$, and $K$ proposed here to represent dichromatic and normal trichromatic vision conform fairly well with the dominator-modulator theory. The $K$ function is the narrowest of the three and since it is unassociated with luminosity could be thought of as a modulator for normal, protanopic, and deuteranopic vision. The $W_p$ function is somewhat narrower than the $W_d$ function and could be thought of as a modulator in the case of normal and tritanopic vision. The spectral location of these functions does not, however, conform perfectly with any of the three preferential regions found by Granit in his study of the eyes of animals.

Granit's view of deuteranopia differs somewhat from that suggested by the functions $W_d$ and $K$. He says,* "Colour-blindness need not, but *can* be possible without parallel change of the photopic luminosity curve. A colour-blindness of this type would be the common form of red-green blindness known as deuteranopia, to be interpreted as absence of the 'red' and 'green' modulators, with the remaining dominator alone giving the normal luminosity curve." The view suggested by the functions $W_d$ and $K$ is that deuteranopia consists of the absence of only the green modulator, $W_p$.

Granit attacks the classical trichromatic theory of color blindness as follows:* "The trichromatic theory regards white as the result of the summed effects of, chiefly, the 'red' and the 'green' sensitivity curves. This forces the theory to accept the consequence that removal of 'red' and/or 'green' should cause removal of the perception of luminosity in the same region of the spectrum. Hence there can be no colour-blindness without profound changes in the form and locus of the luminosity curve. It is an admission of failure to have to explain so important a phenomenon as deuteranopia by pushing it aside to be taken care of by the 'higher centres.' " It is interesting to note that this criticism of classic three-components theory does not apply to the later König form of this theory.

It may be concluded that the functions $W_d$, $W_p$, and $K$ are nearly, if not perfectly, consistent with Granit's dominator-modulator theory, and also that this theory was foreshadowed by König's later views. König, himself, however, began to be interested in the theoretical possibilities of an hypothetical post-receptor center of the visual mechanism apparently first suggested by Donders.[5] He states,† "This is not the place to go into any further discussion of the hypothesis of the existence of a second more centrally located structure of the color system. I wish only to call special attention to the fact that with its assumption many difficulties still existing in the explanation of color vision in the extrafoveal and peripheral parts of the retina, for what I have called the blueblindness of the fovea, for the apparent achromaticity of sensations in the lowest brightness levels, and so forth, would be removed. Through assumption of pathological processes in this more centrally located color apparatus there could be explained further many cases of erythropsia, chloropia and so forth and finally perhaps color sense disturbances due to hysteria and so forth."

This view had been favored for some years by v. Kries‡ and was adopted later by Adams[1] and G. E. Müller.[47,48] Lest it appear that the concise account of dichromatic and normal trichromatic vision afforded by the functions $W_d$, $W_p$, and $K$ establish the correctness of the three-components

* Reference 12, p. 13.

* Reference 12, p. 14.
† Reference 38, p. 730; reference 38a, p. 414.
‡ Reference 39, p. 269.

theory of vision according to König and Granit, and render further theorizing useless, two of the shortcomings of this form of three-components theory may be mentioned, and the way the zone theory of G. E. Müller overcomes these deficiencies may be indicated.

In the long wave end stretch of the spectrum from 700 to 770 mμ the chromaticity is constant; that is, any part of this region of the spectrum can be matched perfectly for a normal observer by any other part provided only that the luminances of the two be equalized. This constancy of chromaticity is indicated in the standard I.C.I. coordinate system by the fact that $X$ and $Y$ bear the same ratio over this wave-length range, and the same thing is indicated from the functions $W_d$, $W_p$, and $K$ from the fact that $W_d$ and $W_p$ also bear a constant ratio over the same range. If, as in König's later view, we regard $W_d$ and $W_p$ as the spectral distributions of the fundamental red and green sensations, the question arises, "What could account for the fact that while varying by a factor of 100 these two distributions still manage to bear exactly the same ratio?" There is only one answer at all plausible; the only way we would expect two processes to bear a constant ratio is that they are both functions of some other single process. Perhaps, for example, $W_d$ and $W_p$ are proportional to excitation of the fibers of the optic nerve in the red and green sense, resulting from receptors (retinal cones, say) containing preponderantly a red-sensitive and a green-sensitive substance, respectively. If the green-sensitive substance were entirely insensitive for wave-lengths greater than 700 mμ, that is, suppose its sensitivity curve corresponded to $W_p$ (see Fig. 4), the explanation for constant chromaticity for wave-lengths longer than 700 mμ would be very simple and believable; that is, only one photosensitive substance, the red sensitive, is affected by radiant energy of this wave-length range. The dependence of the fundamental red sensation upon this kind of radiant energy is customarily ascribed to leakage of a small fairly constant fraction of the green-sensitive substance into the predominantly red-sensitive cones. Thus it is seen that the functions $W_d$, $W_p$, and $K$ though they give a concise account of the connection between normal and dichromatic vision at the level of, perhaps, the fibers of the optic nerve, do not suffice to explain plausibly the constant chromaticity of the long wave end stretch of the spectrum for the normal observer. To provide a satisfactory explanation these functions have to be supplemented by another three-components system referring to the photosensitive substances, themselves.

Although formulation of normal and dichromatic vision in terms of the functions $W_d$, $W_p$, and $K$ gives a clear and satisfactory definition of the samples of radiant energy confused both by an average normal observer and average dichromatic observers, it does not provide any mechanism to explain the usual sensations. That is, these functions describe what the observers confuse but not what they see. Thus for the normal we have the fundamental sensations of red, green, and violet in equal amount summing to neutral. But when the red sensation is missing as in protanopia instead of having green and violet left, we find that blue and yellow are seen. Similarly when the green sensation is missing as in deuteranopia or as in certain extra-foveal portions of the normal retina, we see blue and yellow rather than red and violet as might seem reasonable from the formulation of deuteranopia from $W_d$ and $K$ taken without $W_p$. And finally, when both green and violet are missing, we see neutral colors only as in acquired total color-blindness, rather than the red which might be expected. This failure of the three-components theory to indicate correctly the sensations derivable from the various presumed combinations of the fundamental sensations has been pointed out innumerable times, chiefly by psychologists, who could scarcely be expected to be satisfied with a formulation giving false implications of visual experience even though it did predict correctly which samples of radiant energy are found to be identical in appearance. This failure led to the Hering and Ladd-Franklin theories of vision, and set König, Donders[5] and v. Kries to thinking about higher zones in the color apparatus. This lack is also felt by Granit who remarks,* "The experiments with the cone-eye of the snake suggested that the dominator itself is composed of modulators joined together in such a fashion—either photochemically or by

* Reference 12, p. 13.

connections in the retinal synapses—as to operate as a *functional* unit. This assumption . . . would explain why stimulation of all modulators together also causes an impression of white, and not of all colours confused. . . . Alternatively, the modulators could be coupled in antagonistic pairs which simultaneously neutralized each other at the retinal or some higher level." The first suggestion accords closely with the Adams theory[1,2] which is probably the first to follow explicitly the dominator-modulator form. The latter suggestion corresponds somewhat to the zone theory of G. E. Müller[47] which by 1924 had been elaborated so as to deal satisfactorily in a qualitative way with all known forms of dichromatic and monochromatic vision.

### C. Zone Theory of G. E. Müller

The Müller theory includes three separate color systems, one in the zone of the initial photosensitive substances, a second in the zone of sensory retinal processes aroused by action of the initial photosensitive substances, and a third in the zone of excitations of optic-nerve fibers. The first color system has red, green and violet primaries (R, G, V) and accords with the classical Young-Helmholtz theory. The second system includes two pairs of antagonistic chromatic processes: yellowish red-greenish blue (yR-gB), and greenish yellow-reddish blue (gY-rB). The first named of each pair has a lightening effect; the last named, a darkening effect. And, finally, the third system includes two pairs of antagonistic chromatic excitations and one non-antagonistic pair of achromatic excitations: red-green (r-g), yellow-blue (y-b) and white-black (w-s). White does not cancel black, but combines with it to give gray.

Failure of the yR-gB process corresponds to protanopia. Yellowish red and greenish blue are the hues of the colors confused by the protanope with gray. Since the normal lightening effect of the yR process is missing the theory indicates that the luminosity function of the protanope is shifted toward the short wave end, which corresponds to the facts.

Failure of the gY-rB process corresponds to tritanopia. Greenish yellow and reddish blue are the hues of the colors confused by the tritanope with gray. Since the normal lightening effect of the gY process is missing the theory indicates that the luminosity function of the tritanope is higher than normal at the short wave end. The few data available[7,8,33,34] are inconclusive concerning this indicated deviation from the normal luminosity function.

Failure of the r-g excitation corresponds to deuteranopia. Purplish red and green are the hues of the colors confused by the deuteranope with gray. The theory indicates that the deuteranopic luminosity function corresponds to the normal, which is the same indication given by the functions $W_d$, $W_p$, and $K$ [see Eq. (11)].

If, as seems likely, it is possible to quantify the Müller theory so as to make it embody the same dichromatic confusion colors as those indicated by the functions $W_d$, $W_p$, and $K$, a much more complete account of visual phenomena would be so achieved. The shortcomings of the König-Granit three-components theory outlined above would be overcome. One could then take his choice between a single color system yielding conveniently all of the dichromatic color confusions, or a three-system theory yielding the same confusions together with a prediction of dichromatic sensations.

The major theoretical point of difference between these two views lies in the account of deuteranopia. By the König-Granit theory, deuteranopia and protanopia are retinal defects of similar character. By the Müller theory, deuteranopia is a defect of the optic-nerve fibers; protanopia and tritanopia, defects of the cones. Though there is no conclusive evidence pointing to which, if either, view is correct, it may be significant first that diseases of the nerve-fiber layer of the retinal optic nerve and tract always produce the deuteranopic form of red-green blindness[33] never the protanopic form, and second, there have been several observers found who have tritanopia or protanopia in one eye[4,19,20,21] and normal vision in the other, but never a case of congenital monocular deuteranopia.*

## VIII. SUMMARY

There have been derived, chiefly from the work of Pitt, three functions of wave-length which permit easy definition of the samples of radiant energy found to be identical (metameric)

* Reference 4, p. 78.

not only by the I.C.I. standard observer but also by the two most common types of red-green-blind observers, protanopes and deuteranopes. These three functions $W_d$, $W_p$, and $K$ may be taken together to indicate the metamers of the normal observer; $W_d$ and $K$, taken together indicate the metamers of the average deuteranopic observer; and $W_p$ and $K$, those of the average protanopic observer. The functions $W_d$ and $W_p$ taken together also yield a close approximation to the less well determined metamers of the average tritanopic observer. Methods of using these functions to derive the variation in wave-length of neutral point for dichromatic observers with choice of neutral source are given together with a convenient graphical solution of the ratio of protanopic reflectance to normal reflectance as a function of normal chromaticity. A derivation of the average protanopic, deuteranopic, and tritanopic responses to the Farnsworth dichotomous test for color blindness is also given by way of these three functions. The functions $W_d$, $W_p$, and $K$, therefore, serve to make conveniently accessible the color confusions of dichromatic observers. These three functions also conform to the type of three-components color theory outlined in 1897 by Arthur König, and they are consistent with the dominator-modulator theory proposed by Granit in 1943. They do not, however, form the basis of a complete theory of vision such as that of G. E. Müller.

## REFERENCES


(1) E. Q. Adams, "A theory of color vision," Psychol. Rev. **30**, 56 (1923).

(2) E. Q. Adams, "X–Z planes in the 1931 I.C.I. system of colorimetry," J. Opt. Soc. Am. **32**, 168 (1942).

(3) J. Dalton, "Extraordinary facts relating to the vision of colours," Edinburgh J. Sci. **9**, 97 (1798); reprinted in Edinburgh J. Sci. **5**, 188 (1831).

(4) W. Dieter, "Über die subjectiven Farbenempfindungen bei angeborenen Störungen des Farbensinnes," Zeits. f. Sinnesphysiol. **58**, 73 (1927).

(5) F. C. Donders, "Ueber Farbensysteme," Graefe's Arch. f. Ophthal. [1] **27**, 155–223 (1881).

(6) H. Ebbinghaus, "Theorie des Farbensehens," Zeits. f. Psychol. **5**, 160 (1893).

(7) E. Engelking, "Die Tritanomalie, ein bisher unbekannter Typus anomaler Trichromasie," Arch. Ophthal. Leipsig **116**, 196 (1925).

(8) E. Engelking and H. Hartung, "Spektraluntersuchungen über die Minimalfeldhelligkeiten des Tritanomalen und seine Unterschiedsempfindlichkeit für Änderungen der Helligkeit," Graefe's Arch. Ophthal. **118**, 211 (1927).

(9) Dean Farnsworth, "The Farnsworth-Munsell 100-hue and dichotomous tests for color vision," J. Opt. Soc. Am. **33**, 568 (1943).

(10) Glenn A. Fry, "A quantitative formulation of color mixture and chroma discrimination data for dichromats," J. Opt. Soc. Am. **34**, 159 (1944).

(11) K. S. Gibson and E. P. T. Tyndall, "Visibility of radiant energy," Bur. Stand. Sci. Papers **19**, 131 (1923); S 475.

(12) R. Granit, "A physiological theory of colour perception," Nature **151**, 11 (1943).

(13) J. Guild, "The transformation of trichromatic mixture data: algebraic methods," Trans. Opt. Soc. (London) **26**, 95 (1924–25).

(14) J. Guild, "On the fixed points of a colorimetric system," Trans. Opt. Soc. (London) **32**, 1 (1930–31).

(15) S. Hecht and S. Shlaer, "The color vision of dichromats, I. Wavelength discrimination, brightness distribution, and color mixture, II. Saturation as the basis for wavelength discrimination and color mixture," J. Gen. Physiol. **20**, 57 (1936).

(16) H. v. Helmholtz, *Handbuch der Physiologischen Optik* (Voss, Leipzig, 1885–1896), second edition, p. 370.

(17) H. v. Helmholtz, *Handbuch der Physiologischen Optik* (Voss, Leipzig, 1911), Vol. 2, third edition.

(18) H. v. Helmholtz, *Treatise on Physiological Optics, Translated from the 3rd German Edition*, Edited by J. P. C. Southall (The Optical Society of America, 1924), Vol. 2.

(19) A. v. Hippel, "Ein Fall von einseitiger congenitaler Roth-Grünblindheit bei normalem Farbensinn des anderen Auges," Arch. f. Ophthal. [2] **26**, 176 (1880); [3] **27**, 47 (1881).

(20) Fr. Holmgren, "Ueber die subjective Farbenempfindung der Farbenblinden," Centralbl. f. d. med. Wiss. Berlin **18**, 913 (1880).

(21) Fr. Holmgren, "Fall af ensidig violettblindhet," Upsala läkareförens. förhandl. **16**, 563 (1881); also Centralbl. prakt. Augenh. **5**, 476 (1881).

(22) R. S. Hunter, "Photoelectric tristimulus colorimetry with three filters," Circular of the National Bureau of Standards No. C429 (July 30, 1942).

(23) *Proceedings of the 8th Session, Commission Internationale de l'Éclairage, Cambridge, England* (September, 1931), pp. 19–29.

(24) H. E. Ives, "The transformation of color-mixture equations from one system to another," J. Frank. Inst. **180**, 673 (1915).

(25) D. B. Judd, "Reduction of data on mixture of color stimuli," Bur. Stand. J. Research **4**, 515 (1930); RP163.

(26) D. B. Judd, "The mixture data embodied in the tentative curves of Hecht's theory of vision," J. Opt. Soc. Am. **20**, 647 (1930).

(27) D. B. Judd, "The 1931 I.C.I. standard observer and coordinate system for colorimetry," J. Opt. Soc. Am. **23**, 359 (1933).

(28) D. B. Judd, "Estimation of chromaticity differences and nearest color temperature on the standard 1931 I.C.I. colorimetric coordinate system," J. Research Nat. Bur. Stand. **17**, 771 (1936); RP944; J. Opt. Soc. Am. **26**, 421 (1936).

(29) D. B. Judd, "Specification of color tolerances at the National Bureau of Standards," Am. J. Psychol. **52**, 418 (1939).

(30) D. B. Judd, "Color systems and their interrelation," Illum. Eng. **36**, 336 (March, 1941).

(31) D. B. Judd, "Facts of color-blindness," J. Opt. Soc. Am. **33**, 294 (1943).

(32) K. L. Kelly, K. S. Gibson, and D. Nickerson, "Tristimulus specification of the Munsell *Book of Color* from spectrophotometric measurements," J. Research Nat. Bur. Stand. **31**, 55 (1943); RP 1549; also J. Opt. Soc. Am. **33**, 355 (1943).

(33) H. Köllner, *Die Störungen des Farbensinnes* (Karger, Berlin, 1912).

(34) H. Köllner, "Die Abweichungen des Farbensinnes," in Bethe's *Handb. d. norm. u. pathol. physiol. 12/1, Receptionsorgane II* (Springer, Berlin, 1929), p. 502.

(35) A. König, "Zur Kenntniss dichromatischer Farbensysteme, Wied. Ann. **22**, 567 (1884); Graefe's Arch. Ophthal. [2] **30**, 155 (1884); A. König, *Gesammelte Abhandlungen* (Barth, Leipzig, 1903), p. 11.
(36) A. König, "Die Grundempfindungen und ihre intensitäts-Vertheilung im Spectrum," Sitz. Akad. Wiss. Berlin, p. 805 (July 29, 1886); also *Gesammelte Abhandlungen* (Barth, Leipzig, 1903), p. 60.
(37) A. König and C. Dieterici, "Die Grundempfindungen in normalen und anomalen Farbensystemen und ihre Intensitätsverteilung im Spectrum," Zeits. f. Psychol. **4**, 241 (1893); also A. König, *Gesammelte Abhandlungen* (Barth, Leipzig, 1903), p. 214.
(38) A. König, "Ueber Blaublindheit," Sitz. Akad. Wiss. Berlin, p. 718 (July 8, 1897); (38a) also *Gesammelte Abhandlunen* (Barth, Leipzig, 1903), p. 396; (38b) J. v. Kries, "Ueber Farbensysteme," Zsch. Psychol. u. Physiol. d. Sinnesorg. **13**, 267 (1897).
(39) J. v. Kries, "Die Gesichtsempfindungen," *Nagel's Handbuch der Physiologie des Menschen*" (Braunschweig, 1905), Vol. 3, p. 109.
(40) J. v. Kries, "Note on normal and anomalous colour systems," English edition, *Helmholtz's Treatise on Physiological Optics* (The Optical Society of America, 1924), Vol. 2, p. 402.
(41) C. Ladd-Franklin, "The nature of the color sensations," Appendix to *English Translation of Helmholtz Physiological Optics* (Optical Society of America, 1924), Vol. 2.
(42) E. Ludvigh and E. F. McCarthy, "Absorption of visible light by the refractive media of the human eye," Arch. Ophthal. **20**, 37 (1938).
(43) R. Luther, "Aus dem Gebiet der Farbreiz-Metrik," Zeits. f. tech. Physik **8**, 540 (1927).
(44) J. C. Maxwell, "On the theory of colours in relation to colour-blindness, Letter of Jan. 4, 1855 to G. Wilson," supplements to *Researches on Colour-Blindness*, by G. Wilson (Sutherland-Knox, Edinburgh, 1855).
(45) W. M. McKeon and W. D. Wright, "The characteristics of protanomalous vision," Proc. Phys. Soc. **52**, 464 (1940).
(46) P. Moon, "Proposed standard solar radiation curves for engineering use," J. Frank. Inst. **230**, 613 (1940).
(47) G. E. Müller, *Darstellung und Erklärung der verschiedenen Typen der Farbenblindheit nebst Erörterung der Funktion des Stäbchenapparates sowei des Farbensinns der Bienen und der Fische* (Vandenhoeck & Ruprecht, Göttingen, 1924).
(48) G. E. Müller, "Über die Farbenempfindungen," Zeits. f. Psychol. Ergänzungsb. **17**, **18** (1930).
(49) J. H. Nelson, "Anomalous trichromatism and its relation to normal trichromatism," Proc. Phys. Soc. **50**, 661 (1938).
(50) S. M. Newhall, D. Nickerson, and D. B. Judd, "Final report of the OSA subcommittee on the spacing of the Munsell colors," J. Opt. Soc. Am. **33**, 385 (1943).
(51) W. Ostwald, *Colour Science, Part II, Colour Measurement and Colour Harmony* (authorized translation with an introduction and notes by J. Scott Taylor), (Winsor & Newton, London, 1931), p. 35.
(52) F. H. G. Pitt, "Characteristics of dichromatic vision," Medical Research Council, Report of the Committee on the Physiology of Vision, XIV; Special Report Series, No. 200 (London, 1935).
(53) F. H. G. Pitt, "The nature of normal trichromatic and dichromatic vision," Proc. Roy. Soc. **B132**, 101 (1944).
(54) E. Schrödinger, "Grundlinien einer Theorie der Farbenmetrik im Tagessehen," Ann. d. Physik **63**, 397, 481 (1920).
(55) E. Schrödinger, "Ueber das Verhältnis der Vierfarben-zur Dreifarbentheorie," Sitz. Akad. Wiss. Wien [IIa], **134**, 471 (1925).
(56) L. L. Sloan, "The effect of intensity of light, state of adaptation of the eye, and size of photometric field on the visibility curve,—A Study of the Purkinje phenomenon," Psychol. Monographs **38**, No. 1 (1928).
(57) L. T. Troland, "Report of committee on colorimetry for 1920–21," J. Opt. Soc. Am. and Rev. Sci. Inst. **6**, 527 (1922).
(58) H. V. Walters and W. D. Wright, "The spectral sensivity of the fovea and extrafovea in the Purkinje range," Proc. Roy. Soc. **B131**, 340 (1943).
(59) W. D. Wright, "A re-determination of the trichromatic coefficients of the spectral colours," Trans. Opt. Soc. (London) **30**, 141 (1928–29).

JOURNAL OF THE OPTICAL SOCIETY OF AMERICA VOLUME 35, NUMBER 3 MARCH, 1945

# Spectral Stability in the Condensed Spark Discharge in Air

SAUL LEVY
*Carnegie-Illinois Steel Corporation, Gary Works, Gary, Indiana*

(Received October 20, 1944)

The subject of this paper is the reproducibility of the intensity ratios of spectral lines emitted by a condensed spark discharge in air (denoted here for brevity by spectral stability, or simply, stability). In the first section, the general conditions for such stability are surveyed briefly, and the influence of the circuit parameters is discussed. In the second section, the function of auxiliary gaps in various schemes is considered. In the last section, measurements with an electrical set-up suitable for the purposes of quantitative spectrum analysis are given.

## I.

THE question of stability arises because of the practical impossibility of exactly reproducing all the working conditions. Since some variation of these conditions is unavoidable, the resulting instability must be kept within reasonable limits. The following factors influence the intensity ratios of the emitted lines of different elements: the evaporation of the substances in the electrodes at high temperatures, the relative

excitation of different atomic energy levels, and the re-absorption. The chosen working conditions determine what influence accidental variations of each of these factors have on the emitted radiation.

The relative excitation of different energy levels is very sensitive to temperature fluctuations at lower temperatures of the discharge. Because of approximately exponential distribution of the atoms among the energy levels in a

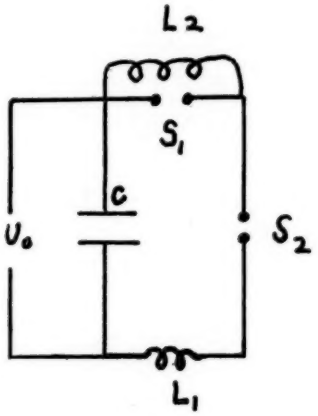



FIG. 1.

discharge at atmospheric pressure, not necessarily corresponding to the same temperature in different parts of the discharge across and along the spark channel, the fluctuations of the relative excitation with variations of the temperature decrease with increasing temperature. The relative excitation then approaches a constant value, equal to the ratio of the statistical weights of the corresponding energy levels. An example of the relative influence of temperature fluctuations on the stability at different temperatures is given in an earlier paper.[1] In general, the change of the relative excitation for two levels with energy difference $\Delta\epsilon$ is proportional to $(\Delta\epsilon \cdot dT/T^2) \exp(-\Delta\epsilon/kT)$, where $dT$ is the change of temperature and $k$ is the Boltzmann constant. The time for approaching equilibrium between different energy levels at atmospheric pressure is of the order of $10^{-6}$ sec., which is about one-tenth of the period of oscillation in the usual circuits. Since the temperature increases with the current density, it is possible to obtain high temperatures in the discharge with relatively small average currents, using current impulses of large momentary density.

Another advantage of this type of excitation is that the heating of the electrodes is prevented. The heating of solid electrodes is a relatively slow process, in comparison with the time for establishing a state of equilibrium in gases and vapors. When the discharge is only of short duration, the electrodes are not heated to any considerable extent. The vapor density is then relatively small. This condition prevents the lowering of the effective temperature (which would otherwise occur because of the low excitation and ionization potentials of metallic vapors), and diminishes and stabilizes reabsorption.

Thus, by using currents of high density and of short duration, one obtains high temperatures of the gases and vapors in the discharge gap, with relatively low temperatures of the electrodes. This is essential for stable intensity ratios of spectral lines.

The influence of the circuit parameters on the current density is therefore of interest, as the latter determines the effective temperature, as well as the duration of the discharge. In the well-known approximative expression for the current $i$ in a discharge in a circuit with *pure ohmic* resistance $R$, self-inductance $L$, and capacitance $C$, charged to the voltage $V_0$, namely,

$$i \simeq V_0 C^{\frac{1}{2}} L^{-\frac{1}{2}} \cdot \exp(-R/2L \cdot t) \sin t/(CL)^{\frac{1}{2}}, \quad (1)$$

the first factor represents the initial current amplitude, the second factor, the damping which determines the duration of the current. This latter factor is modified in the discharge through a spark gap[2] in that the damping factor is expressed in the form $e^{-\delta t}$ where $\delta$ is the sum of the pure ohmic damping $\delta_0$ and that because of the spark $\delta_s$. $\delta_s$ is not constant at different stages of the discharge and depends on the initial amplitude, the ohmic damping, the frequency, and the material of the electrodes. The resulting decrease of the amplitude is larger than in (1) and approaches in many cases a linear one.

[1] S. Levy, J. App. Phys. **11**, 483 (1940).

[2] D. Rojansky, Ann. d. Physik **36**, 281 (1911); *Electric Rays* (Russian) (Richter, 1916). J. Zenneck and H. Rukop, *Lehrbuch der drahtlosen Telegraphie*, (Hirzel, 1925). See also the review by S. M. Raiskij in J. Tech. Phys. USSR, **10**, 431 (1940).

Clearly this damping does not affect the average current, but determines the duration of the discharge. The equivalent resistance of the spark in circuits with parameters similar to those now in use in spectrographic analysis, has been found to be of the order of 0.05 ohm.[3] This makes it understandable why defective contacts in the circuits have such an adverse effect on stability; they affect mostly the damping, when the ohmic resistance of the circuit is low.

Obviously, the average current cannot be characteristic of the resulting temperature of the vapors during the emission, if the averaging is extended over the whole period of time between two complete discharges of the condenser, because the actual discharge is usually of very short duration. An average taken over the time $\tau=1/\delta$, corresponding to a decrease of the current to $e^{-1}$, gives a more representative value of the effective current. In every case the current is proportional to the initial amplitude $V_0C^{\frac{1}{2}}L^{-\frac{1}{2}}$. This amplitude may reach very substantial values. With $C=0.005$ $\mu$f, $L=0.02$ mh and $V_0=15{,}000$ volts the expression gives over 200 amp. Considering the very narrow discharge channel, of the order of 0.1-mm diameter,[4] one obtains for the current density values of the order of magnitude of $5.10^5$ amp./cm$^2$.[5]

Although the stability of excitation increases with the temperature, there is a limiting temperature beyond which it is impracticable to go, because of intensity of the "background" (continues and band spectra) relative to the intensity of metallic lines, and also because of excessive line broadening.

## II.

Although it is simple, by choice of proper parameters of the circuit, particularly, by decreasing the self-inductance, to obtain the desirable high momentary current density, even with a relatively small power supply, maintaining the proper sparking voltage requires some care. Since the current density is proportional to the charging voltage, the reproducibility of the voltage and its absolute value are essential. Whatever the rating of the transformer is, the condenser cannot be charged to a voltage higher than the sparking potential of the gap. The simple expedient of using only one gap (between the samples), which is usually relatively small, does not permit charging the condenser to a high enough voltage. Therefore the application of an auxiliary gap is necessary. In the Feussner scheme there is a rotary gap, operated by means of a synchronous motor that can be adjusted to let the discharge pass at the maximum voltage. Using a stationary auxiliary gap,[6] the lengths of both the analytical and the auxiliary gaps must be such as to charge the condenser to the required voltage. (Regarding prevention of the passage of several discharges, if not desired,

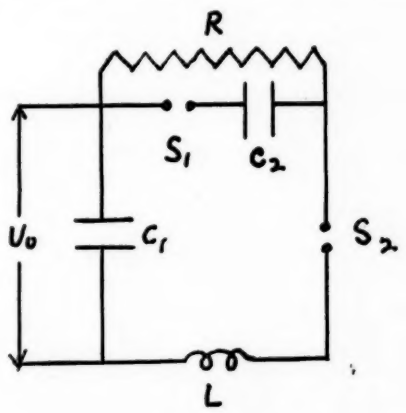



FIG. 2.

during each half-period of the primary frequency, see Twyman.[6] A suitably chosen resistance between the secondary of the transformer and the condenser also regulates the number of sparks.) Both schemes have the following feature in common: The sparking voltage, which depends

[3] Originally found by Braun, confirmed by K. Simons, Ann. d. Physik **13**, 1044 (1904); D. Rojansky (reference 2) and others.

[4] J. Slepian, Elec. World **91**, 761 (1928), theoretically. S. M. Raiskij [J. Tech. Phys. USSR **10**, 529 (1940)]; experimentally, using rotating mirror.

[5] During the discharge the ions diffuse away from the original channel and one could expect that this would result in a gradual broadening of the discharge channel for oscillations with a small damping and in a lowering of the current density. However, according to L. Loeb, *Fundamental Processes of Electric Discharges in Gases* (John Wiley & Sons, Inc., New York, 1939), p. 175, the diffusion of ions is negligible at atmospheric pressure during a time of the order of $10^{-3}$ sec.

[6] F. Twyman, *Spectrochemical Analysis of Metals and Alloys* (Chemical Publishing Company, Brooklyn, New York, 1941), p. 129; W. Zehden, J. Soc. Chem. Ind. **59**, 236 (1940).

on the gap length and on the shape and material of the electrodes, is subject to fluctuations because of occasional variations of these factors in both the analytical and the auxiliary gaps.

Maresca[7] and Simons[8] introduced a scheme (Fig. 1), for measuring the spark resistance, in which the charging voltage on the condenser $C$, that is, the sparking voltage of the auxiliary gap $S_2$, does not depend upon the gap $S_1$ at all. During the charging period of the condenser the gap $S_1$ is shorted by the coil $L_2$. The length of the gap $S_1$, and the shape and material of its electrodes have no influence on the sparking voltage which is determined by the auxiliary gap $S_2$, only. However, as soon as the sparking voltage is reached and oscillations in the circuit $C-L_2-S_2-L_1-C$ start, the gap $S_1$ breaks down because of the high potential difference on the ends of the coil $L_2$. The length of the gap $S_1$ cannot be larger than that of $S_2$, of course, but can be otherwise of any length. The current continues to oscillate through these gaps in series, avoiding the coil $L_2$ for the duration of the oscillations. Thus, the coil $L_2$ and the spark gap $S_1$ eliminate one another alternatively. The current density and damping do not depend appreciably on small variations of the gap length $S_1$ when $S_2$ is large, as the resistance of an ionized gap is approximately proportional to the length of the gap. The gap $S_1$ has no influence on the discharged energy.

TABLE I. Electrodes: 2 steel pins.

| Percent Si | Ni | Cr | Mo |
|---|---|---|---|
| $0.19_5$ | 0.83 | $0.76_5$ | $0.18_6$ |
| $0.20_2$ | 0.82 | $0.77_5$ | $0.18_1$ |
| $0.20_9$ | 0.81 | 0.77 | $0.18_8$ |
| $0.19_9$ | 0.83 | 0.75 | $0.18_8$ |
| $0.20_0$ | 0.82 | 0.74 | $0.18_6$ |
| $0.19_8$ | $0.81_5$ | 0.75 | $0.18_6$ |
| $0.19_8$ | 0.83 | 0.77 | $0.18_6$ |
| $0.20_0$ | 0.83 | 0.78 | $0.19_1$ |
| $0.21_0$ | 0.83 | 0.76 | $0.18_7$ |
| $0.19_1$ | 0.82 | 0.76 | $0.19_1$ |
| $0.19_2$ | 0.82 | 0.76 | $0.18_7$ |

The lines used in this investigation were:

| | | | |
|---|---|---|---|
| Si I | 2881.6 | Fe I | 2918.0 |
| Ni I | 3012 | Fe I | 3011.5 |
| Cr II | 2876 | Fe II | 2876.8 |
| Mo II | 2816 | Fe II | 2827.4 |

[7] A. Maresca, Physik. Zeits. **14**, 9 (1902). Maresca used a high resistance instead of a coil. The coil was introduced by Simons.

Raiskij[8] recently applied this scheme in his investigation of the relative emission in different parts of the spark gap of variable length. It was of importance in his investigation that the current density remain as constant as possible. For brevity this scheme will be designated in the following by M-S-R (the Maresca-Simons-Raiskij scheme).

A modification of the M-S-R scheme is shown in Fig. 2, where $R$ is a high resistance or a suitable inductance coil. The charging capacitance in this scheme is different from the capacitance of the oscillating circuit. The passage of the low frequency current from the secondary of the transformer through the analytical gap is prevented.

## III.

The conditions on the surface and the shape of the electrodes of the gap $S_1$ may influence the local heating and the evaporation or sputtering of the electrodes, thus affecting the resulting intensities of spectral lines in the same way as in other schemes mentioned above. It can be, nevertheless, expected that the independence of charging the condenser of the analytical gap in the M-S-R scheme, or in its above-described modification, will render it a suitable light source for quantitative spectrum analysis. In order to test the scheme, a small power sparking unit was assembled consisting of a 0.25-kva transformer with 15,000 v effective in the secondary, which charged a condenser of 0.005 $\mu$f. An inductance of 0.02 mh was added to the oscillating circuit having originally approximately 0.005 mh. The analytical gap of 5-mm length was shorted with a coil of 1 mh. As is to be expected, large variations of the inductance of this coil produced changes neither in the absolute brightness nor in the intensity ratios of the spectral lines. The auxiliary gap, with electrodes made of copper rods of 7 mm $\phi$,* was 17 mm long. No care was needed in maintaining the shape of the auxiliary electrodes, the surface of which became slightly curved after some sparking and retained this shape permanently. The electrodes of the analytical gap were either 2 steel pencils of $\frac{1}{4}''$ diameter or pellets made of steel

[8] S. M. Raiskij, J. Tech. Phys. USSR **9**, 1719 (1939).
* See also reference 7.

drillings, of the same diameter, and sparked against a pointed graphite rod. The surfaces of the samples were flat. A 30 seconds exposure after 10 seconds presparking was applied.[9] The resulting blackenings of the spectrograms made with a large quartz prism spectrograph, without a condensing lens, slit to source distance 50 cm, were suitable for analysis of the usual constituents of steel (Si, Mn, Cu, Ni, Cr, Mo, V, Al, Ti, Cb, Zr and some other elements), starting with the lowest concentrations of importance for control and check analysis, as a rule with 0.01 percent or lower.

Extensive series of spectrograms were prepared with this source. Several thousand determinations were made for various elements, usually in groups of 10 spectrograms of the same samples on each plate in order to eliminate the error caused by plate calibration. The photometric error, consisting of the error of the reading with a commercial densitometer as well as of a possible non-uniformity of the developed plate, are included in the (arithmetical) average error given below. The densitometer slit used was $0.01 \times 1$ mm$^2$, the spectrographic slit 0.02 mm. Repeated measurements of the same plates showed that the average error of the densitometer reading and concentration calculation was $\pm 0.8$ percent. The measurements were made under the conditions of a busy industrial laboratory and the results can be considered representative. As is to be expected, the reproducibility of the intensity ratios (or concentration determinations) depended largely upon the selection of the line pairs and, perhaps, to some extent, on the element. Occasional substantial deviations from the average values, of the order of 5 percent, could mostly be traced to unsatisfactory contacts in the circuits including the contacts between the holders and the samples.

Typical series of the results of repeated sparking of a sample are reproduced in Table I.

The reproducibility of the results vary somewhat with the elements. This must not, however, be considered necessarily as an indication of a different degree of uniformity of distribution of different elements in the samples, or as a result of different volatility of the elements. Selection of another working pair often improves the reproducibility of the determination of the concentration in an element. An example is the analysis of copper in steel. In a series of measurements, using Cu I 3274 and Fe I 3739.4, the average deviation was $\pm 2.5$ percent, whereas with the lines Cu I 3274 and Fe I 3286.8 it was $\pm 2$ percent and with Cu II 2242.6 and Fe II 2255.17 only $\pm 1.6$ percent.

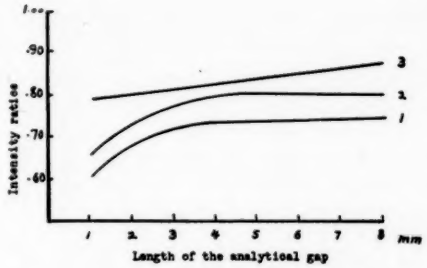



FIG. 3. Intensity ratios of spectral lines in spark of changeable length. 1. Cr. II 2860.9—Fe II 2827.4. 2. Fe II 3096.3—Fe I 3098.2 3. Ni I 3012–Fe I 3011.5.

The average deviation from the mean in a short series of measurements as given in Table I usually shows only the order of magnitude of the deviation. The deviations for Ni, using the above-mentioned lines, from very numerous measurements, averaged $\pm 1.5$ percent, whereas for Cr the deviation was $\pm 1.2$ percent. Subtracting quadratically the error of the densitometer readings and of the concentration computation ($\pm 0.8$ percent) we obtain for the deviations of Cr determinations, caused by the source, approximately $\pm 0.9$ percent. The reproducibility with pellets was inferior to that with solid pins. For Cr the total average deviation, using pellets, was approximately $\pm 2$ percent.

The material of the auxiliary electrodes did not seem to be of importance for the reproducibility of the results (copper, silver, tungsten, and graphite electrodes were tried). Variations of the length of the auxiliary gap from 17 mm down to about 12 mm did not change the intensity ratios of investigated lines noticeably. This relatively small sensitivity to changes of the (large) auxiliary gap may perhaps be explained by the assumption that sufficiently high temperatures of the discharge were obtained. (See Part I.)

In all the measurements described above the

[9] This presparking was of very little influence on the results [see also F. G. Barker, J. Iron Steel Inst., 139, 211 (1939)]. This does not apply to more powerful units with a larger capacitance.

analytical gap was spaced using a caliper of 5-mm thickness. Results obtained with a changeable analytical gap are given in Fig. 3. The ordinates are intensity ratios of spectral lines, the abscissas the distances between the electrodes (2 steel pencils). It is interesting to note that the usually very changeable intensity ratio of the two Fe (arc and spark) lines does not change noticeably with the increase of the gap when the latter is over 3 mm. A variation of 1 mm of the gap, when it is larger than 3 mm, is practically not reflected in the intensity ratios of all lines investigated. These results have, naturally, only a relative significance, as the intensity ratios depend on the optical arrangement, with apertures largely influencing the effective relative radiation from different parts of the discharge. The difference in the behavior of the three line pairs with the increase of gap is believed to be connected with the different effect of the heating of the electrodes, as a short spark heats the electrodes less than a longer one does. No air blast was applied for cooling the electrodes in any of the experiments.

**SUMMARY**

With fixed conditions of evaporation or sputtering of the electrodes, the spectral stability of the condensed spark discharge depends largely upon the effective temperature of gases and vapors. Fluctuations of the temperature resulting from accidental variations of the working conditions, affect relatively little the intensity ratios of the spectral lines at high temperatures of the discharge, if the electrodes are not heated up considerably. Both conditions are usually fulfilled when high current densities of a short duration are used.

One of the causes of instability of the discharge, namely, the dependence of the discharged energy upon the analytical gap, is eliminated in an electrical scheme with a stationary auxiliary gap in series and with a coil or a resistance in parallel with the analytical gap (Maresca-Simons-Raiskij). In this scheme the electrodes of the auxiliary gap, which can be kept easily at standard conditions, determine the discharged energy. Experimental testing of this scheme shows its suitability for the purposes of quantitative spectrum analysis.

The help of the members of the Gary Works Spectrographic Laboratory in making the numerous tests for this investigation is gratefully acknowledged.

---

JOURNAL OF THE OPTICAL SOCIETY OF AMERICA VOLUME 35, NUMBER 3 MARCH, 1945

# The Application of Spectrochemical Analysis in the Steel Mill*

P. R. IRISH
*Development and Research Department, Bethlehem Steel Company, Bethlehem, Pennsylvania*
(Received December 26, 1944)

IN the manufacture of alloy steel it is necessary to analyze the molten metal a number of times during the refining period. One of the factors contributing to the total heat time is the time required to make these analyses. Spectrochemical methods can be used to advantage for this analytical control, because of their greater speed, which leads to increased production and closer control of the composition. At the Alloy Division of the Bethlehem, Pennsylvania Plant of Bethlehem Steel Company, the spectrographic method has been used since March, 1941. Here the spectrographic laboratory presently makes about 8000 separate determinations a week, covering many elements, over wide ranges. We believe it to be the most extensive application of spectrochemical methods to a direct composition control problem and the first installation to apply such control to alloy steel.

In Fig. 1 is a chart showing the demand on the laboratory in a typical twenty-four hour period.

* Paper presented at the Twenty-Ninth Annual Meeting of Optical Society of America; October 20 and 21, 1944, New York, New York.

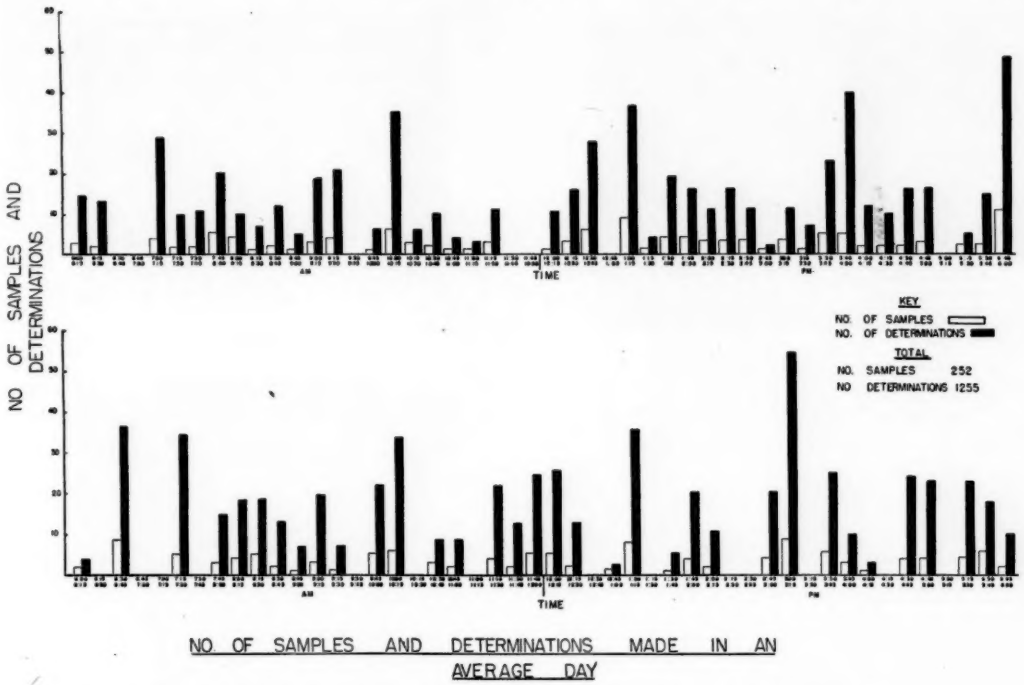



FIG. 1. Chart showing number of samples and determinations made in an average day.

On the chart are indicated the number of samples received and the number of determinations required for each fifteen minute period during the twenty-four hours. The maximum demand occurred during the period 3:00–3:15 A.M., when the laboratory received 9 samples requiring a total of 55 separate determinations. To make this large volume of control analyses rapidly with one set of spectrographic equipment it is necessary to use six operators on each shift, who, in the course of twenty-four hours, expose, develop, and read 180 to 190 plates containing about 1000 spectrograms. Two of the operators are stationed in the source room (Fig. 2). The sample electrodes which consist of two $\frac{7}{32}''$ diameter pins, $1\frac{1}{2}''$ long, arrive in this room by pneumatic tube and are prepared for analysis by one operator who grinds a domed surface on one end of each electrode. The other operator loads the specimens into the electrode clamps and makes all of the necessary exposures. Depending on the analyses required, the specimen electrodes are excited with the spark source, the arc source, or both. Also, spark spectra may be made in two spectral regions. For concentrations above 1.50 percent, duplicate spark spectra are made in one region. Spark exposures are 40 seconds with a prespark of 20 seconds and arc exposures are 30 seconds with no preburn.

FIG. 2. Source room.

One operator handles the work in the darkroom. Photographic plates are developed by the rapid processing technique.[1] However, the solu-

[1] H. B. Vincent and R. A. Sawyer, Spectrochem. Acta [2] I, 131 (1939).

FIG. 3. Microphotometer room.

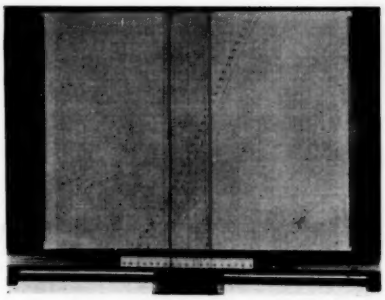


FIG. 4. Calculating board.

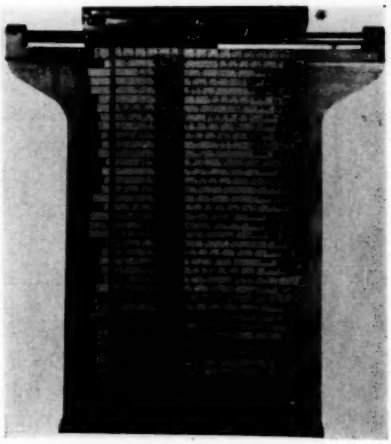


FIG. 5. Scale board.

tions are agitated in a mechanism that imparts two-directional agitation to the solutions. The two motions operate out of step with each other, each having a different length of stroke and oscillating period. This results in an irregular turbulent motion of the solution and prevents the set wave pattern obtained with rocker type agitation generally used in developing spectrographic plates.

A view of the microphotometer room is shown in Fig. 3. Three operators are stationed in this room. One reads the microphotometer, the second transfers the density readings into percent concentration, and the third operator sends the results to the electric furnaces or the open hearth shops by means of the telautograph. The second operator computes the percentage values with the use of two boards. One board (Fig. 4), containing a plot of the plate-characteristic curve, is used for transposing the galvanometer readings for the iron reference line and element line into the logarithm of the intensity ratio for these two darknesses. The hair line of the scale board (Fig. 5) is then set on the lower scale to the same logarithmic ratio of intensity value and the percentage concentration is read from the scale above which corresponds to the lines read on the microphotometer. The photographic plate characteristic curve is made by reading the blackening of a series of iron lines of different densities and plotting the galvanometer readings against corresponding logarithmic intensity values which have been previously established.

With this procedure analyses of samples taken during the refining of a heat are reported in less than fifteen minutes. A single determination is usually made in nine minutes, although, during peak load conditions, analyses for eight elements may take as long as twenty-five minutes.

Because of the speed and large number of analyses required, covering many elements extending over wide ranges, all determinations of concentration must be based on working curves which have been previously established. Quantitative spectrographic analysis would not be practical for large-volume high speed analysis if it were necessary to expose on each plate a group of standards or even one standard in order to obtain the required accuracy. Therefore, successful results depend on maintaining the calibration of the equipment over an extended period. I should like to discuss some of the factors that must be considered if rapid analyses are to be made and an acceptable accuracy obtained.

One of the major problems of rapid routine spectrochemical analysis is the control of the photographic process. Many methods of calibrating the photographic plate have been devised, but none has proven entirely satisfactory, especially in the calibration of the toe and shoulder of the curve. Since the present procedure does not give a strictly accurate plate calibration, steps must be taken to minimize such errors.

The chart in Fig. 6 shows the ranges and elements for which the equipment is calibrated. A list of the spectral line pairs employed is given in Table I. More line pairs are used than would seem necessary. However, by employing a large number of pairs regardless of the concentration, the spread in density between the iron reference line and the element line will never be large. By keeping this difference small no serious error is introduced, even though the plate calibration curve is incorrect. All of the analytical curves for the element nickel are shown in Fig. 7. The entire extent of each curve indicates the range that can be covered by the photographic plate (Spectrum Analysis No. 1) while the heavy sections indicate the portion that is used in routine analytical control. As shown, the working curves are used only over a short section of their total length. In the lower concentrations, where a larger percentage error is tolerable, the working curves can be safely used over a wider range.

A further restriction on the line pairs is necessary since it is not possible to draw the ends of the plate characteristic curve accurately. The blackening readings for all lines must fall within narrow limits if consistent results are to be obtained. In Fig. 8 is a plate characteristic curve as it is plotted in our laboratory, in the wavelength range 2500A to 2600A. The solid section indicates the portion used in analytical control

TABLE I. Spectral line pairs.

| (1) Element line | (2) Iron line | (3) Range percent | (4) Freq. cal. is checked | (5) Source | (6) Filter |
|---|---|---|---|---|---|
| Mn 2933.06 | Fe 2931.60 | 0.05 –0.20 | 1 mo. | Spark | |
| Mn 2933.06 | Fe 2936.91 | 0.20 –0.45 | 1 mo. | Spark | F |
| Mn 2933.06 | Fe 2897.26 | 0.45 –0.65 | 1 mo. | Spark | F |
| Mn 2558.59 | Fe 2553.73 | 0.65 –1.50 | ½ mo. | Spark | F |
| Si 2881.58 | Fe 2885.35 | 0.002–0.015 | 6 mo. | Arc | |
| Si 2881.58 | Fe 2886.32 | 0.015–0.06 | 2 mo. | Arc | |
| Si 2881.58 | Fe 2875.30 | 0.06 –0.10 | 2 mo. | Arc | F |
| Si 2881.58 | Fe 2874.17 | 0.10 –0.70 | 1 mo. | Spark | |
| Ni 3414.77 | Fe 3431.81 | 0.01 –0.05 | 6 mo. | Arc | |
| Ni 2316.04 | Fe 2318.35 | 0.05 –0.25 | 2 mo. | Spark | |
| Ni 2316.04 | Fe 2322.33 | 0.20 –0.50 | 2 mo. | Spark | F |
| Ni 2303.00 | Fe 2307.31 | 0.50 –1.00 | 8 hrs. | Spark | F |
| Ni 2300.09 | Fe 2294.61 | 1.00 –1.60 | 8 hrs. | Spark | |
| Ni 2302.47 | Fe 2303.35 | 1.60 –2.40 | 8 hrs. | Spark | |
| Ni 2302.47 | Fe 2304.73 | 2.40 –3.50 | 8 hrs. | Spark | F |
| Ni 2302.47 | Fe 2307.31 | 3.50 –5.50 | 8 hrs. | Spark | F |
| Cr 3593.49 | Fe 3612.07 | 0.005–0.04 | 6 mo. | Arc | |
| Cr 2677.16 | Fe 2676.88 | 0.03 –0.12 | 2 mo. | Spark | |
| Cr 2677.16 | Fe 2682.52 | 0.12 –0.25 | 2 mo. | Spark | F |
| Cr 2677.16 | Fe 2642.02 | 0.25 –0.45 | 2 mo. | Spark | F |
| Cr 2862.57 | Fe 2876.80 | 0.45 –1.20 | 1 mo. | Spark | F |
| Cr 2818.36 | Fe 2837.29 | 1.20 –1.90 | 1 mo. | Spark | |
| V 3185.40 | Fe 3207.09 | 0.005–0.02 | 6 mo. | Arc | |
| V 3110.71 | Fe 3106.56 | 0.02 –0.10 | 1 mo. | Spark | |
| V 3110.71 | Fe 3075.72 | 0.10 –0.40 | 1 mo. | Spark | F |
| V 3130.27 | Fe 3075.72 | 0.40 –1.50 | ½ mo. | Spark | F |
| Mo 3170.35 | Fe 3207.09 | 0.002–0.008 | 6 mo. | Arc | |
| Mo 3170.35 | Fe 3196.13 | 0.008–0.03 | 1 mo. | Arc | F |
| Mo 2816.15 | Fe 2815.51 | 0.03 –0.15 | 1 mo. | Arc | |
| Mo 2775.40 | Fe 2776.18 | 0.15 –0.55 | 1 mo. | Spark | |
| Mo 2775.40 | Fe 2770.51 | 0.55 –2.00 | 1 mo. | Spark | F |
| Mo 2644.35 | Fe 2682.52 | 2.00 –3.50 | 1 mo. | Spark | F |
| Mo 2644.35 | Fe 2649.46 | 3.50 –6.50 | 1 mo. | Spark | F |
| Cu 3273.96 | Fe 3277.35 | 0.04 –0.20 | 1 mo. | Spark | |
| Cu 3273.96 | Fe 3286.76 | 0.20 –0.50 | 1 mo. | Spark | F |
| Sn 3175.02 | Fe 3157.04 | 0.003–0.01 | 6 mo. | Arc | |
| Sn 3034.12 | Fe 3012.45 | 0.01 –0.03 | 1 mo. | Arc | |
| Sn 2839.99 | Fe 2848.72 | 0.03 –0.10 | 1 mo. | Arc | F |
| Sn 3330.59 | Fe 3331.61 | 0.10 –0.20 | 1 mo. | Arc | |
| Al 3082.16 | Fe 3080.11 | 0.004–0.02 | 2 mo. | Arc | |
| Al 3082.16 | Fe 3078.43 | 0.02 –0.05 | 2 mo. | Arc | F |
| Al 3082.16 | Fe 3053.44 | 0.05 –0.12 | 2 mo. | Arc | F |
| Al 3082.16 | Fe 3075.72 | 0.80 –2.00 | 1 mo. | Spark | |
| W 4008.75 | Fe 3996.97 | 0.05 –0.10 | 2 mo. | Arc | |
| W 4008.75 | Fe 4016.43 | 0.10 –0.25 | 2 mo. | Arc | |
| W 2656.54 | Fe 2651.71 | 0.25 –0.50 | 2 mo. | Arc | |
| W 2656.54 | Fe 2656.15 | 0.50 –1.00 | 2 mo. | Arc | |
| W 2397.09 | Fe 2415.06 | 2.00 –5.00 | * | Spark | F |
| Ti 3371.45 | Fe 3389.75 | 0.004–0.04 | 2 mo. | Arc | |
| Ti 3371.45 | Fe 3387.41 | 0.04 –0.07 | 2 mo. | Arc | F |
| B 2496.78 | Fe 2505.01 | 0.002–0.008 | * | Arc | |
| B 2496.78 | Fe 2496.99 | 0.008–0.05 | * | Arc | F |
| Co 2414.46 | Fe 2422.68 | 0.03 –0.30 | * | Arc | |

* Curve used only occasionally, checked each time.

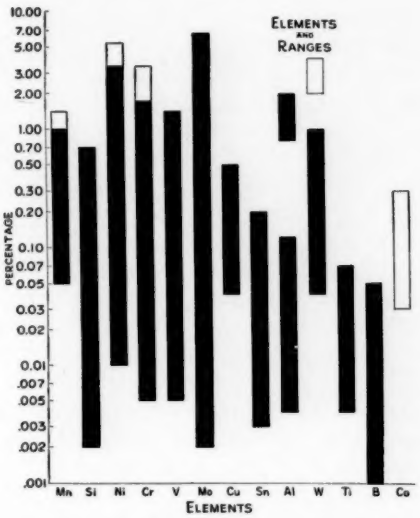



FIG. 6. Chart showing elements and ranges.

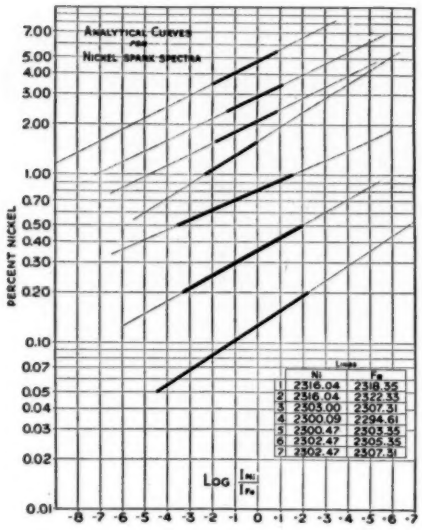



FIG. 7. Analytical curves for the element nickel.

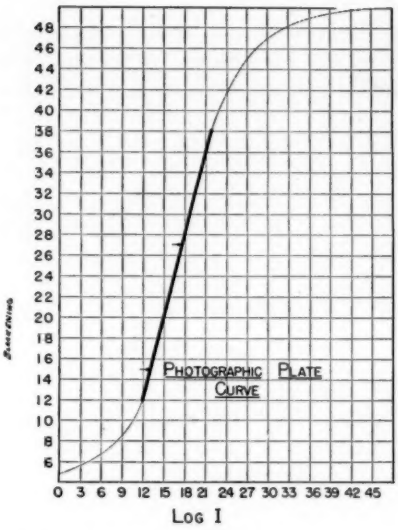



FIG. 8. Photographic plate characteristic curve.

for lines above 2400A. When determining the nickel concentration with the lines in the 2300A to 2325A range, the readings are held to even closer limits, as indicated by the arrows. Narrower limits are required because of the low contrast of the plate at this wave-length. The number of lines that will have densities in these narrow ranges has been increased by using a filter directly in front of the slit of the spectrograph. This filter reduces the intensity of the light entering one-half of the slit by 35 percent. Line pairs read in the filtered section of the spectrum are indicated in column 6 of Table I.

The nature of the photographic plate brings about a further restriction on the selection of the line pairs, in addition to the requirement that the two lines must have about the same density. This requirement is, that the two lines must have approximately the same wave-length. If this is not the case serious error will be introduced, for two reasons. An error will be introduced because the plate does not have the same sensitivity and contrast at different wave-lengths. Second, points more widely separated suffer more from random non-uniformities of the photographic plate than do points close together. When a reliable light source is used, these factors over-shadow any errors that may arise because of inconsistency in the relative intensity of the two lines, introduced by the fact that the element line may originate at a lower energy level than the reference line. A greater precision will be obtained if line pairs close together in wave-length are used, although according to theory a line, say 25A away, should give better consistency.

Contrary to some beliefs, tests show that the largest factor contributing to the total error in large-volume, rapid spectrographic analysis is the photographic plate error. Kaiser[2] and Vincent and Sawyer[3] concluded that the photographic error was small. However, Kaiser, in connection with his study of the factors that contribute to the total spectrochemical error, eliminated all but the random photographic error by confining all measurements to spectrograms exposed on one plate, including those for making the working curve. Vincent and Sawyer greatly reduced the chances for photographic error by limiting their study to a few plates presumably all exposed in a period of a few hours. Such studies are not at all indicative of the factors influencing spectrochemical precision in a laboratory exposing as

[2] H. Kaiser, Zeits. f. tech. Physik [7] **17**, 227 (1936).
[3] H. B. Vincent and R. A. Sawyer, Jr., J. Opt. Soc. Am. **32**, 686 (1942).

many as 7000 spectrograms per week, spread over 1300 photographic plates.

Thus, in large volume analysis, the greatest concern is not the precision shown by a series of analyses read from a group of spectrograms exposed on one plate nor with a group exposed on successive plates; but the deviation shown by analyses obtained from spectrograms that are made several months apart, during which period perhaps 20,000 plates have been used, containing a total 100,000 separate spectra. Here, not only the randon photographic error is of importance but also the error caused by inconsistent contrast-wave-length characteristics, variations in sensitivity-wave-length characteristics, and failure of the plate calibration method to interpret all densities correctly. Thus when these factors are taken into account the photographic error and thus the total spectrographic error will be larger than that indicated by either of studies mentioned above.

In routine rapid spectrochemical analysis light source errors contribute less to the total error than the photographic variations in the ranges of high concentrations where an interrupted spark is employed. This source is supplemented with an a.c. arc for use in the lower ranges, which are below the sensitivity of the spark source. When employing an a.c. arc, source errors contribute more to the total error than any other factor. However, we are more concerned with the interrupted spark precision, since this source is used in the high range of concentration where a greater degree of accuracy is required. The adjustment of the electrical factors of the spark source has been largely empirical. These factors must be set to give the best accuracy which may not necessarily be the condition of best precision for any one of the several analyses to be made. Thus the adjustments must be such that only one set of working curves will be required for all of the different alloys covered by the ranges of Fig. 6. If a different set of curves were required for each alloy range, the procedure would become too cumbersome and slow. Thus, for example, the manganese content of the steel must be read from the same curve whether the steel contains no alloy or as much as 5.00 percent to 8.00 percent alloy. Also, the heat treatment of the steel should have no effect on the calibration. A set of conditions which we have found satisfactory are a condenser voltage of 40,000 volts, a capacity of 0.016 microfarad, and a damping resistance of 3 ohms with the circuit containing no added inductance. These factors are for a commercial condenser having a mica dielectric. The conditions have been designed to burn the samples without visible fusion and with a high enough power to overcome the effects of heat treatment and the influence of the varying compositions existing between the different grades of steel. These conditions would have to be altered if a different size or shape electrode were selected as the standard size.

Considerable attention has been given to the ruggedness of the design of the spark source. Our circuit is powered by a 15 kva oil-cooled transformer, the condenser is made up of 16 Type 462–52, 0.004 microfarad, Cornell Dubilier mica capacitors. Two of these units are used in series to form a section. Eight sections in parallel make up the complete condenser bank. All electrical connections are made with $\frac{1}{8}'' \times 1''$ copper buss. It is essential that the electrical parts be heavy if they are to maintain their electrical rating and thus not cause variations in the working-curve calibrations because of shifts in the spark circuit itself. The entire source including the a.c. arc is assembled in the enclosure shown in Fig. 2. A sturdy construction has been provided by making the optical bench with the electrode clamps an integral part of the structure. The electrode clamps are located in the box mounted on the two horizontal rails. The electrical connections to the clamps are made by means of arms which extend out the back from the center of the box. These arms pass through the two horizontal slots between the two tracks and make contact with two rails mounted on the back of the panel. It is these two back rails, extending the length of the bench, that carry the high voltage. As the electrode clamps are moved along the bench the arms continue to make contact with the rails by means of a sliding connection. Since the high voltage rails are inaccessible the operators are not endangered by exposure to open cable connections. Furthermore, all doors to the controls as well as to the electrode clamps are equipped with trap switches, thus completely protecting the operators from elec-

trical shock. The controls are conveniently arranged, including the timers for controlling the length of the exposures and the shutter of the spectrograph. Arc and spark exposures are made with the same electrode clamps. The transfer from one source to the other is done by means of one of the knife switches.

In rapid spectrographic analysis an effort is made to use lines that can be covered by a single setting of the spectrograph in order to save time and avoid out-of-focus spectrograms that might be encountered if the region is changed. To increase the range of a single setting, instruments have been designed that will handle plates longer than ten inches. However, in our Bethlehem Plant Laboratory a Bausch and Lomb large Littrow instrument is employed, and a single setting of this spectrograph will not cover all of the lines given in Table I which extend from 2294–4009A. Therefore, it was necessary to devise a system for changing the spectral region rapidly and accurately. This was done by changing the spectrograph to a two-region instrument. The focus and plate tilt are set automatically for the two regions by means of pneumatic cylinders which actuate the plate head and prism table between stops. The region can be changed in less than two seconds. In an installation now under construction spark spectra will be automatically exposed in two regions after the source has been placed in operation. Part way through the exposure the shutter of the spectrograph will close, the plate will rack, and the spectral region will change. The shutter will then open and the remainder of the exposure will be made in the other region. Under present operating conditions the spectral region is changed about 4,000 times each week.

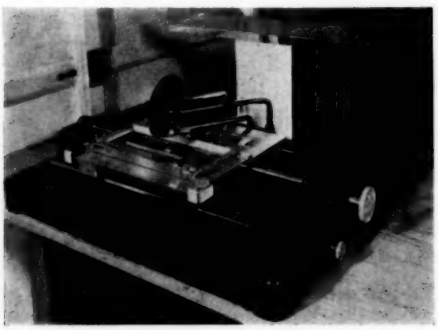


FIG. 9. Microphotometer.

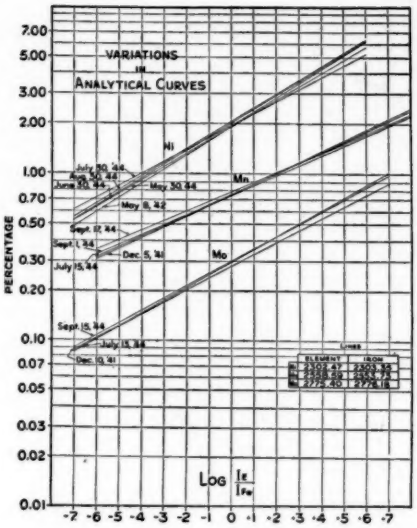



FIG. 10. Variation in the analytical curve for nickel, manganese, and molybdenum.

The microphotometer employed is shown in Fig. 9. This instrument was designed by Vincent and Sawyer[4] especially for rapid routine spectrochemical analysis. The features of this instrument are ease of operation and ruggedness of construction, which lead to reliable reading even with severe use in the hands of many operators. Once a plate has been placed in the instrument the operator can make all readings with little shifting of position. While reading, his left hand rests on the front left-hand corner of the plate carriage. With this hand he can move the main carriage rapidly along, shifting any portion of the spectrum on the screen. At the same time his left hand is in position for shifting any one of the spectra on the screen by moving the cross carriage. The operator's right arm rests on the table with his hand extending to the large reading knob. To align the microphotometer slit with the spectral lines, the operator only has to move his right hand to the small knob in front of the driving knob. An image of the unknown spectrum is projected on the viewing screen with a master spectrum below. This screen and the density

[4] H. B. Vincent and R. A. Sawyer, Jr., J. Opt. Soc. Am. **31**, 639 (1941).

scale are so arranged that practically no adjustment in the accommodation of the eye is required when the gaze of the operator is shifted from one to the other.

Under the conditions described, most of the analytical curves maintain their calibration over many months. In fact some of the curves, especially those in the lower concentration ranges where the highest precision is not required, have not been altered since the equipment was originally calibrated in December, 1941. Maintaining the calibration curves for the element nickel in the range above 1.00 percent has proved the most difficult. These curves are checked every eight hours and in the course of four weeks it might be found necessary to make an adjustment in the calibration. The variation found in one nickel working curve over a period of four months is shown in Fig. 10. Calibration variations for two other curves are also shown in Fig. 10. The frequency with which all curves are tested with standards is given in column 4 of Table I.

Under the conditions described here it has been possible to maintain a precision of a standard deviation of 2 percent to 3 percent of the concentration for spark spectra. An error of this magnitude has proven acceptable for alloy ranges up to 3 percent and 4 percent. Spectrum lines located in a favorable position on the plate give greater precision. Tests made at concentrations where the difference in density between the element line and reference line is small also show a greater degree of precision. For certain elements in complex alloys, in ranges up to 8 percent, the spectrographic precision is found to be better than that of chemical methods. However, an improvement in the precision is desirable if rapid spectrochemical analysis methods are to be extended to the concentration ranges of 10 percent and 20 percent with entirely satisfactory results.

---

JOURNAL OF THE OPTICAL SOCIETY OF AMERICA VOLUME 35, NUMBER 3 MARCH, 1945

# The Visual Effect of Non-Uniform Surrounds*

PARRY MOON, *Massachusetts Institute of Technology, Cambridge, Massachusetts*

AND

DOMINA EBERLE SPENCER, *Tufts College, Medford, Massachusetts*

## 1. INTRODUCTION

A PREVIOUS paper[1] has correlated data on minimum perceptible contrast and minimum visual angle as functions of helios. In this earlier work, the visual field was uniform except for the test object; and numerous researches show that such uniformity gives best conditions for seeing. But uniform visual fields are not encountered in practice. Thus it becomes advisable to consider how the previous results are altered by non-uniformity in the helios distribution. Non-uniformity may be caused by one or more concentrated glare sources or it may consist of a helios distribution that is a continuous function of position.

The purpose of the present paper is twofold: first to investigate the visual effect of a non-uniform field, and second to establish criteria on helios uniformity, for use in lighting design. Cone vision alone is considered. In the investigation of criteria, three conditions are used, leading to slightly different recommendations. These three conditions are: (a) Eyes fixed on the work, non-uniform surround. (b) Eyes adapted to the brightest or darkest part of the field. The effect of suddenly looking back at the work is determined. (c) Eyes adapted to the work. Transient shock caused by suddenly looking from work to any part of the field is evaluated.

All these and many other conditions may

* Based on a paper presented at the 29th Annual Meeting of the Optical Society of America, New York, October 21, 1944.

[1] P. Moon and D. E. Spencer, J. Opt. Soc. Am. **34**, 605 (1944).

obtain in practice. It is felt, however, that if reasonable agreement can be found in the criteria established on these three bases, such criteria should be of practical value in the design of lighting for good vision.

In the investigation of the effect of non-uniformity, extensive use is made of the previous work on adaptation.[2] Holladay's principle[3] allows a great simplification, since all possible helios distributions that give the same adaptation helios may be considered as equivalent. For condition (a), these ideas allow equations and curves to be formulated to give a correction factor that can be applied to the visual data obtained with a uniform surround. The same principles can be applied to (b). The conclusion is that the adaptation helios obtained by staring at the brightest part of the field should not exceed 3 times the adaptation helios for the work. Also, the adaptation helios for the darkest part of the field should not be less than $\frac{1}{3}$ that for the work. For (c) transient analysis on the basis of photochemical theory gives a similar criterion but one that varies greatly with adaptation helios.

## 2. HISTORY

An extensive literature is available on non-uniform visual fields and on the practical applications of the results to problems in automobile headlighting, street lighting, and room lighting. The subject is usually called "glare," and great ingenuity is exercised in breaking this subject into subdivisions characterized by the adjectives "veiling," "brightness," "blinding," "dazzle," "disability," "discomfort," "contrast," "spot," and "scotomatic." Since no one has ever been able to define these words precisely, it seems better to omit them and to talk about *non-uniform fields*.

The *visual field* may be defined, for a given position of the observer's eyes, as the portion of space that sends light to the retina. By a *uniform* field is meant a visual field whose helios is the same throughout except for a possible test spot of small size. All other fields are said to be non-uniform.

Researches on non-uniform fields may be divided into two classes: (1) Tests with a fixed line of sight and with a concentrated source at an angle $\theta$ from the line of sight. The effect of the source is evaluated by measuring minimum perceptible contrast with and without the glare source. (2) Tests in which the observer looks directly at a spot of high helios and evaluates it subjectively.

In the first group belong a large number of researches that show that vision is less satisfactory when a concentrated source is introduced into a uniform visual field. For example, in 1900 Uhthoff and Depène[4] investigated the effect of a concentrated source at various angles $\theta$ from the line of sight. They found that minimum perceptible visual angle $\alpha$ increased as the angle $\theta$ was reduced. Other researches, such as that of Bordoni,[5] dealt with non-uniformity arising in street lighting and other situations found in practice.

An investigation of fundamental importance is that of Holladay,[3] who first established a mathematical relation for the effect of a concentrated source. This result was corroborated by Stiles,[6,7] who extended and applied the ideas of Holladay. The experiments of these investigators dealt with minimum perceptible contrast as affected by sources in the field of view. Similar tests were made by Cobb and Moss[8] and by Lythgoe,[9] using visual angle instead of contrast. Most of the work was confined to a single concentrated source superposed on a uniform field. Lythgoe, however, made a few tests with surrounds which were uniform but which had a different helios from that of the background. He found that dark surrounds were deleterious as well as bright surrounds.

In the second type of test, neither contrast nor visual angle is employed as a criterion of goodness of seeing. The observer adapts to a large uniform surround of helios $H_A$. A spot of higher helios is exposed momentarily, and the observer decides

[2] P. Moon and D. E. Spencer, J. Opt. Soc. Am. **33**, 444 (1943).
[3] L. L. Holladay, J. Opt. Soc. Am. **12**, 271 (1926).
[4] W. Uhthoff, Arch. f. Ophth. **32**, 171 (1886); **36**, 33 (1890); R. Depène, Monats. Augenhlk, **38**, 289, 390 (1900).
[5] U. Bordoni, Elettrotecnica, **11**, 585 (1924).
[6] W. S. Stiles, Proc. Roy. Soc. **B104**, 322 (1929).
[7] W. S. Stiles and C. Dunbar, Illum. Res. Comm. (Great Britain), Tech. Paper No. 16, 1935.
[8] P. W. Cobb and F. K. Moss, Trans. I.E.S. **23**, 1104 (1928).
[9] R. J. Lythgoe, Report of Comm. Physiol. of Vision, London, No. 10 (1932).

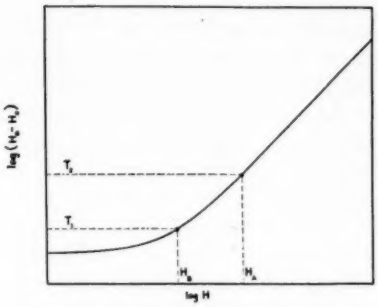


FIG. 1. The helios threshold $(H_B - H_0)$ as a function of helios. Adaptation to a uniform field of helios $H_B$ gives the threshold $T_1$. A glare source introduced into this uniform field raises the adaptation helios to $H_A$ and increases the threshold to $T_2$.

whether the shock caused by the sudden appearance of the spot is unpleasant. It would be expected that the precision obtainable in such a test would be much lower than in Type (1). However, several experimenters have obtained data that are in surprisingly good agreement.

Nutting[10] investigated the effect of adaptation helios and developed an empirical formula for objectionable helios of the spot, a formula that has had wide acceptance by illuminating engineers. Holladay[3] included the effect of size. Summaries are given by Millar and Gray[11] and by Fowler and Crouch.[12] Numerous attempts[13] have been made to apply these data to lighting design, though such applications have been limited by the lack of a generalized theory and by the limited range of variables used in the experiments.

### 3. THE HOLLADAY PRINCIPLE

In a previous paper,[1] data were given on minimum perceptible contrast and on minimum visual angle for foveal vision. In all the tests, the field was of uniform helios. But in lighting practice, the field is non-uniform; and thus it is necessary to determine how much reduction in vision results from a given non-uniformity in the visual field.

The classical research on the subject was made by Holladay.[14] As a result of his experiments, he developed an idea which may be called the Holladay principle: *If any non-uniform visual field producing an adaptation helios $H_A \geqq H_B$ is replaced by a uniform field of helios $H = H_A$, the visual thresholds will be unaltered.*

The transition from the original experiments to the Holladay principle in its most general form involves a number of steps. The reasoning behind these steps is somewhat involved and has never been clearly stated. Thus it may be advisable to consider the procedure by which the Holladay principle was developed. The visual field will be divided into two parts:[2] (a) A central part which is imaged on the fovea. This part will be taken as a circular disk whose radius subtends a visual angle of 0.0131 radian (diameter 1.5°). The circular disk will ordinarily contain a test object having a helios $H_0$. The remainder of the disk will be called the *background* and will have a uniform helios $H_B$. (b) The *surround*, consisting of all the visual field outside the circular disk (*a*). The surround may contain a glare source subtending a solid angle $\Delta\omega$ and having a helios $H_G$. Notation will be as follows: $H_S$ = helios (blondel) of uniform surround; $H(\theta, \varphi)$ = helios of surround, a function of $\theta$ and $\varphi$; $\theta$ = angle (radian) between source and angle of sight; $\varphi$ = angle measured about line of sight (see reference 2); $H_B$ = helios of background; $H_0$ = helios of test object; $H_A$ = adaptation helios.

The 5 steps taken in the formulation of the general Holladay principle will now be considered.

(1) The first step consists in obtaining a curve of the threshold $(H_B - H_0)$ as a function of $H_S$. The entire visual field is uniform $(H_S = H_B)$ except for a circular spot subtending a visual angle of perhaps 1°. The helios $H_0$ of the center spot is varied until it can just be distinguished

[10] P. G. Nutting, Trans. I.E.S. **11**, 939 (1916).
[11] P. S. Millar and S. McK. Gray, Proc. C.I.E., 277 (1928).
[12] E. W. Fowler and C. L. Crouch, Illum. Eng. **36**, 897 (1941).
[13] E. E. Potter and P. Meaker, Trans. I.E.S. **26**, 1025 (1931). E. E. Potter and W. G. Darley, Trans. I.E.S. **35**, 775 (1940). A. B. O'Day and W. Sturrock, Trans. I.E.S. **31**, 351 (1936).

[14] Reference 3. The first mention of the Holladay principle seems to have been made in a paper by M. Luckiesh and L. L. Holladay, "Glare and visibility," presented before the Annual Convention of the Illuminating Engineering Society in October, 1924. This paper was published in the Trans. I.E.S. **20**, 221 (1925). It contained no data, however, and the curves showed no experimental points. A more satisfactory presentation was made by Holladay in 1926 (reference 3). Further work of a similar nature is given in L. L. Holladay, J. Opt. Soc. Am. **14**, 1 (1927).

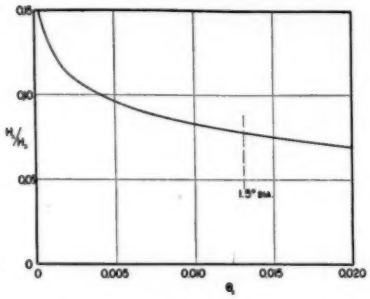



FIG. 2. Contribution of the surround. A large uniform surround of helios $H_s$ is varied in size by changing the angle $\Theta_1$ subtended by the internal radius. As $\Theta_1$ approaches 0, the adaptation helios would approach $\infty$ according to Eq. (2). For the recommended value of $\Theta_1=0.0131$ radian, however, the entire surround contributes only 7.74 percent of the total adaptation helios.

from its surroundings of helios $H_B$. If the test spot is exposed only momentarily, the eye will remain adapted to the background, or $H_A=H_B=H_S$. Thus the threshold $(H_B-H_0)$ can be plotted against $H_A$ and the result will be a curve like that of Fig. 1.

Figure 1 applies to cone vision[15] but a similar curve is obtained for rod vision. At low values of $H_A$, the curve has a horizontal portion. At sufficiently high values of $H_A$, the threshold increases linearly with helios. At intermediate values, there is a gradual transition from one extreme case to the other. Such curves have been obtained by a number of investigators[1] and the form of curve is well established.

(2) The next step is to introduce a concentrated glare source into the uniform field. Suppose that the position of the glare source is fixed ($\theta=$const., $\varphi=$const.) and that all the rest of the field is of uniform helios $H_S=H_B$. The experiment consists in varying $H_G$ through a wide range, keeping all other factors constant. When this is done, a curve is obtained of $(H_B-H_0)$ *vs.* $H_G$. *This curve is found to have the same shape as that obtained in Fig. 1.* In fact, if definite ordinates are chosen arbitrarily and the corresponding values of $H_A$ and $H_G$ are read from the two curves, a plot of $H_A$ *vs.* $H_G$ is found to be a straight line, or

$$H_A=H_S+kH_G, \qquad (1)$$

where $k$ is the slope of the line. Therefore, a uniform field with one superposed glare source can be replaced by a completely uniform field, provided the helios of the latter is adjusted in accordance with Eq. (1). Thresholds will then be the same in the two cases and Fig. 1 is used with either.

The experiment may be performed in other ways. Evidently $H_S$ may be varied with $H_G$ constant. Or the size of the glare source may be altered or the angle $\theta$ changed. In all these cases, an equivalent adaptation helios can be calculated according to the Holladay principle. The experimentally determined relation is

$$H_A=H_S+\frac{KH_G}{\pi\theta^2}\cos\theta\,\Delta\omega, \qquad (2)$$

where $K=9.6\times10^{-3}$ if the values of helios are in blondels, the angles in radians, and the solid angle is steradians.

Equation (2) was found by Holladay[3] and was confirmed by the thorough researches of Stiles,[6] and Stiles and Crawford.[15] The range of the variables employed in the investigations is shown in Table I. The helios of the surround was varied from 0 to 15,000 blondels, the angle $\theta$ from 1° to 30°, and the glare flux density on a plane at the eye and perpendicular to the line of sight $(D_A=H_G\cos\theta\cdot\Delta\omega/\pi)$ ranged from 0 to 500

TABLE I. Investigations of the Holladay principle.

| Authority | Range of $H_S$ (blondel) | $\frac{H_G\Delta\omega}{\pi}\cos\theta$ (lumen m$^{-2}$) | $\theta$ (deg.) |
|---|---|---|---|
| Luckiesh and Holladay, 1925 | 1 to 1000 | 0 to 100 | 5 to 30 |
| Holladay, 1926 | 1 to 100 | 0 to 200 | —— |
| Holladay, 1927 | $2\times10^{-4}$ to 30 | 0 to 6 | 2.5 to 25 |
| Stiles, 1929 | 0 to 30 | 0 to 10 | 1 to 10 |
| Stiles and Crawford, 1934 | 0 to 15,000 | 0 to 500 | —— |

TABLE II. Effect of changing $\Theta_1$. $\Theta_2=1.0$, $Ci(2\Theta_2)=0.4230$

| Diam. of background | $2\Theta_1$ (radian) | $Ci(2\Theta_1)$ | Contribution of surround, $(H_A/H_S)$ | Contribution of background, $(1-H_A/H_S)$ |
|---|---|---|---|---|
| 0.0057° | 0.0001 | −8.6331 | 0.184 | 0.816 |
| 0.057 | 0.001 | −6.3305 | 0.140 | 0.860 |
| 0.29 | 0.005 | −4.7211 | 0.109 | 0.891 |
| 0.75 | 0.0131 | −3.7580 | 0.091 | 0.909 |
| 1.0 | 0.0175 | −3.4684 | 0.085 | 0.915 |
| 1.5 | 0.0262 | −3.0650 | 0.077 | 0.923 |
| 2.0 | 0.0349 | −2.7784 | 0.072 | 0.928 |
| 2.3 | 0.0400 | −2.6421 | 0.069 | 0.931 |

[15] W. S. Stiles and B. H. Crawford, Proc. Roy. Soc. **B116**, 55 (1934); Phys. Soc., London, "Discussion on vision," 194 (1932).

lumens m$^{-2}$. The effect of a glare source is almost negligible beyond 30°, so the lack of data beyond that point is of no consequence. The helios range covers all problems of street lighting and interior lighting and most problems of daylighting. Thus the Holladay principle is firmly established for a single glare source on a uniform field and may be employed as a foundation for studies in adaptation and glare.

(3) The effect of *position* of the glare source is now investigated. Holladay[3] found that the angle $\varphi$, which fixes the position of the glare source about the line of sight, has no appreciable effect. Thus the problem has axial symmetry, as might be expected from the approximate symmetry of the eye. This conclusion was checked by Stiles.[6] Holladay found also that several glare sources at the same angle $\theta$ were replaceable by a single source having the same total $H_G\Delta\omega$. Thus for $n$ glare sources placed anywhere in the field, Eq. (2) can be replaced by

$$H_A=H_S+\frac{K}{\pi}\sum_{i=1}^{n}\frac{(H_G)_i}{\theta_i^2}\cos\theta_i\cdot\Delta\omega_i. \qquad (3)$$

The subject of additivity of glare sources has been investigated also by Crawford.[16]

(4) All of the experiments were made with concentrated glare sources superposed on a uniform field. A special case occurs when $H_S=0$. Then, according to Eq. (3),

$$H_A=\frac{K}{\pi}\sum_{i=1}^{n}(H_G)_i\theta_i^{-2}\cos\theta_i\cdot\Delta\omega_i. \qquad (3a)$$

But any continuous helios distribution may be considered as made up of a combination of small contiguous glare sources. Therefore, any helios distribution $H(\theta, \varphi)$ produces the adaptation helios,

$$\begin{aligned}H_A&=\frac{K}{\pi}\int_{\omega}H(\theta,\varphi)\cdot\theta^{-2}\cos\theta\,d\omega\\&=\frac{K}{\pi}\int_{\Phi_1}^{\Phi_2}d\varphi\int_{\Theta_1}^{\Theta_2}H(\theta,\varphi)\cdot\theta^{-2}\sin\theta\cos\theta\,d\theta,\end{aligned} \qquad (4)$$

where the boundaries of the visual field are denoted by $\Phi_1$, $\Phi_2$, $\Theta_1$, $\Theta_2$. This extension of the Holladay principle was made in a previous paper.[2] Its validity depends on the truth of the Holladay principle and the additivity principle (3). No other assumptions have been made, so Eq. (4) is just as true as the experimentally determined principles on which it is based.

(5) An examination of Table I shows that the only place where data are lacking is at $\theta<1°$. Obviously, none of the foregoing equations will hold as $\theta\to0$ for they give infinite values at zero

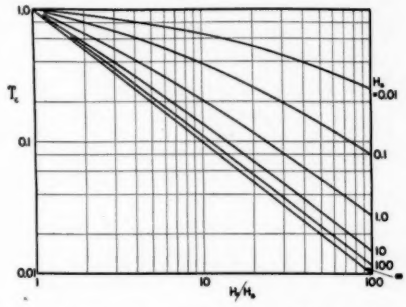



FIG. 3. Delos ($T_c=c_B/c_A$) as a function of background helios $H_B$ and adaptation helios $H_A$.

angle. There is no reason to expect that the effect of a concentrated glare source would be worse when imaged on the center of the fovea than anywhere else in the fovea. It is well known that the fovea contains no rods and that the cones are packed tightly together and form a homogeneous population out to $\theta=0.75°$. Thus the $\theta^{-2}$ relation should hold up to the boundary of the fovea, but at smaller angles the effect should be independent of angle.

To make Eq. (4) useful, therefore, one must assume a definite value of $\theta$, below which Eqs. (1) to (4) do not apply. We will now determine a value of $\Theta_1$ to separate the background from the surround. With axial symmetry and a surround of uniform helios $H_S$, Eq. (4) becomes

$$H_A/H_S=0.0192\left[1-\frac{\sin 2\Theta_2}{2\Theta_2}-Ci(2\Theta_1)+Ci(2\Theta_2)\right], \qquad (5)$$

where $Ci$ is the cosine integral. This equation gives the contribution to $H_A$ made by all the surround of helios $H_S$ from angle $\Theta_1$ to angle $\Theta_2$. It has been shown[2] that the outer limit has very

[16] B. H. Crawford, Proc. Phys. Soc. London, **48**, 35 (1936).

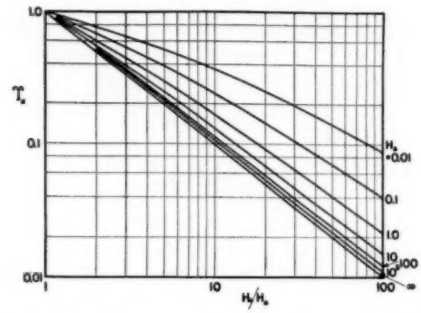



FIG. 4. Delos ($\tau_\alpha = \alpha_B/\alpha_A$) as a function of $H_B$ and $H_A$.

little effect and may be chosen as $\Theta_2 = 1.0$ radian. The effect of changing the inner boundary is then calculated from Eq. (5). The results are shown in Fig. 2. Even if the uniform surround is carried down to $\frac{1}{4}°$ from the line of sight, $H_A/H_S$ is only 0.10 or the entire surround contributes only $\frac{1}{10}$ of the total. The remainder of the foveal adaptation $(1-H_A/H_S)$ must be contributed by the small disk of radius $\frac{1}{4}°$.

The calculated values for various assumed diameters are given in Table II. The fovea is certainly not smaller than 1° diameter nor greater than 2°, and this range causes a difference of only one percent in $(1-H_A/H_S)$. Thus any reasonable size may be taken without appreciable change. We have chosen[2] a diameter of 0.0262 radian (1.5°) as the boundary between the fovea and the parafovea and the boundary between the surround and the background.

Equation (4) can then be modified to express the adaptation helios caused by the entire visual field:

$$H_A = 0.923 H_B + \frac{K}{\pi}\int_{\Phi_1}^{\Phi_2} d\varphi \int_{\Theta_1}^{\Theta_2} H_s(\theta, \varphi) \cdot \theta^{-2} \sin\theta \cos\theta \, d\theta. \quad (4a)$$

If the field has axial symmetry,

$$H_A = 0.923 H_B + 0.0192 \int_{0.0131}^{1} H_S(\theta) \cdot \theta^{-2} \sin\theta \cos\theta \, d\theta. \quad (4b)$$

These equations reduce to Eqs. (1)–(3) for the special cases there considered and they allow the calculations of $H_A$ for all possible helios distributions. The most surprising result of this analysis is that *for a uniform field, the portion imaged on the fovea contributes over 90 percent of the total foveal adaptation.* The recent experimental work of Crawford[17] affords an approximate verification of this conclusion. Crawford varied a uniform surround from 3° diameter to 60° diameter and found no detectable difference in adaptation over this range.

Equations (4), (4a), and (4b) allow the calculation of adaptation helios for any kind of visual field. Data and equations[1] are available also for the minimum perceptible contrast $c_{\min}$ and the minimum visual angle $\alpha$, both obtained with uniform surrounds. Holladay's principle makes possible a tremendous expansion in the usefulness of these visual data. No longer are they limited to the special case of uniform surrounds, for Holladay's principle says that they apply to any helios distribution, provided that the adaptation helios $H_A$ can be calculated for the distribution. And Eq. (4) provides the means for this calculation.

## 4. THRESHOLD, CONTRAST, AND VISUAL ANGLE FOR $H_A \geqq H_B$

As shown in Section 3, the experimental evidence proves that the same curve of threshold

TABLE III. Lythgoe's experimental results on the effect of a dark surround.

| $H_B$ (blondel) | Obs. | Visual angle $\alpha$ (radian) $H_S = 0$ | $H_S \cong H_B$* | Ratio |
|---|---|---|---|---|
| Lythgoe, Table V, p. 67 | | | | |
| 135 | T. B. | $146\times10^{-6}$ | $136\times10^{-6}$ | 0.93 |
| 135 | J. M. | 165 | 148 | 0.90 |
| 135 | J. M. | 171 | 150 | 0.88 |
| 135 | S. M. | 166 | 145 | 0.87 |
| 135 | S. M. | 156 | 145 | 0.93 |
| 135 | Mean | 161 | 145 | 0.90 |
| Lythgoe, Table VI, p. 69 | | | | |
| 0.55 | J. M. | 462 | 403 | 0.87 |
| 0.55 | S. M. | 339 | 313 | 0.92 |
| 4.6 | J. M. | 257 | 227 | 0.88 |
| 4.6 | S. M. | 226 | 226 | 1.00 |
| 135 | J. M. | 168 | 157 | 0.93 |
| 135 | S. M. | 161 | 148 | 0.92 |
| 420 | J. M. | 171 | 145 | 0.85 |
| 420 | S. M. | 160 | 141 | 0.88 |
| | Mean | | | 0.90 |

* Optimum values. In some cases, lowest visual angle was obtained when $H_S$ was somewhat less than $H_B$ rather than at exact equality.

[17] B. H. Crawford, Proc. Roy. Soc. **B129**, 94 (1940).

$(H_B-H_0)$ *vs.* $H_A$ is applicable for a uniform field with any combination of superposed glare sources, the $H_A$ being calculated in each case by means of Eq. (3). For a completely uniform field, the threshold can be plotted against $H_A$ and the result is a curve like that of Fig. 1. As shown in a previous paper,[1] the data on cone vision can be expressed by the equation,

$$(H_B-H_0)_{\min}=c_\infty[A_1+H_A^{\frac{1}{2}}]^2, \qquad (6)$$

where $c_\infty$ and $A_1$ are constants:

$$c_\infty=0.0123, \quad A_1=0.808.$$

The same curve applies when glare sources are superposed on the uniform field. In Fig. 1 the threshold rises from $T_1$ with the uniform field to $T_2$ when a glare source is added. As might be expected, the threshold is increased by the introduction of the glare source, which means that vision is inferior with a glare source superposed on a uniform field. Note that we consider here only cases where the adaptation helios is greater than (or equal to) the background helios.

Equations can be written also for *contrast*. As noted previously,[1] contrast is defined as

$$c=(H_B-H_0)/H_B. \qquad (7)$$

Equation (6) can be expressed in terms of contrast by dividing both sides by $H_B$. Thus for the special case of a uniform field, $H_A=H_B$ and

$$c_B=(H_B-H_0)/H_B=\frac{c_\infty}{H_B}[A_1+H_B^{\frac{1}{2}}]^2. \qquad (8)$$

For the non-uniform field, with background still at $H_B$ but with adaptation to $H_A$, Eqs. (6) and (7) give

$$c_A=\frac{c_\infty}{H_B}[A_1+H_A^{\frac{1}{2}}]^2. \qquad (9)$$

It is convenient to employ the ratio of the two contrasts, or

$$\frac{c_B}{c_A}=\left(\frac{A_1+H_B^{\frac{1}{2}}}{A_1+H_A^{\frac{1}{2}}}\right)^2=\left[\frac{K_1+1}{K_1+(H_A/H_B)^{\frac{1}{2}}}\right]^2, \qquad H_A\geqq H_B, \quad (10)$$

where $K_1=A_1H_B^{-\frac{1}{2}}$. The ratio $c_B/c_A=T_c$ can be plotted against $H_A/H_B$, as shown in Fig. 3. The quantity $T_c$ is a sort of "relative visibility" and is called *delos*,[1] a Greek word meaning "visible." It never exceeds unity, and its value indicates how much reduction in vision occurs because of the introduction of the glare source.

For each value of background helios, $K_1$ is a constant, and a curve of $T_c$ *vs.* $H_A/H_B$ can be computed. Figure 3 shows that such curves are practically coincident for $H_B$ above 100 blondels. Fortunately, the values of helios encountered in room lighting are frequently above this limit, in which case Eq. (10) assume a particularly simple form. If $H_B>100$, $K_1\ll 1$ and Eq. (10) becomes

$$T_c=1/(H_A/H_B), \quad H_A\geqq H_B. \qquad (11)$$

The equation for *minimum visual angle* $\alpha$ is of the same form[1] as the equation for minimum perceptible contrast $c_{\min}$. Thus it seems reasonable to develop an equation for $\alpha$ similar to Eq. (9). The minimum visual angle may be written as

$$\alpha_B=\frac{1}{H_B}f(H_B), \qquad (12)$$

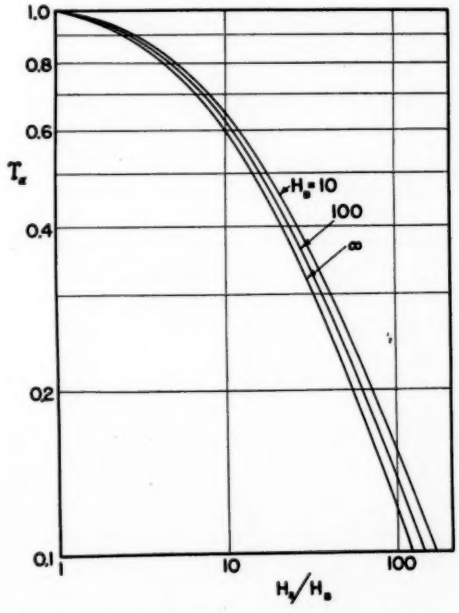



FIG. 5. Delos for a uniform surround of helios $H_s$. The curves show how the relative visual acuity is reduced as the helios of the surround is raised above the helios of the background. $\Theta_1=0.0131$ radian, $\Theta_2=1.00$ radian.

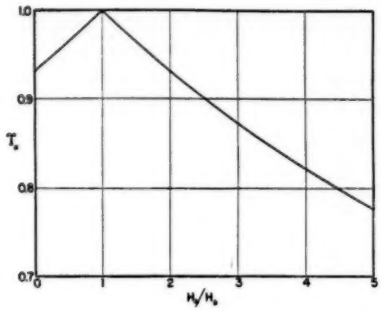



FIG. 6. Delos for a uniform surround of helios $H_s$. Here $H_B=100$ blondels, $\Theta_1=0.0131$ radian. Curve is same as the middle curve of Fig. 5 but is extended to ratios $H_s/H_B$ less than 1.0.

when $H_A=H_B$. In analogy with Eqs. (8) and (9), let us write

$$\alpha_A=\frac{1}{H_B}f(H_A), \tag{13}$$

when $H_A \geqq H_B$. This leads to—*Postulate I: When the adaptation helios $H_A$ is greater than (or equal to) the background helios $H_B$, the minimum visual angle is given by Eq. (13).*

From the previous paper,[1]

$$f(H_A)=\alpha_\infty[A_2+H_A^{\frac{1}{2}}]^3, \tag{14}$$

where $\alpha_\infty=118.3\times10^{-6}$, $A_2=0.412$. Thus, by Postulate I,

$$\alpha_A=\frac{118.3\times10^{-6}}{H_B}[0.412+H_A^{\frac{1}{2}}]^3, \quad H_A\geqq H_B. \tag{15}$$

Analogous to Eq. (10), the expression for *delos with respect to visual angle* is

$$\Upsilon_\alpha=\frac{\alpha_B}{\alpha_A}=\left[\frac{K_2+1}{K_2+(H_A/H_B)^{\frac{1}{2}}}\right]^3, \quad H_A\geqq H_B, \tag{16}$$

where $K_2=A_2H_B^{-\frac{1}{2}}$. Curves are plotted in Fig. 4. The general slope is similar to the curves of Fig. 3; and at sufficiently high helios, the curve becomes identical to that of Fig. 3. In both cases if $H_B>100$, the curve is a straight line with slope $-1$ on a log-log plot. Thus, for $H_B>100$,

$$\Upsilon_\alpha=\Upsilon_c=1/(H_A/H_B), \quad H_A\geqq H_B. \tag{17}$$

It would be advisable to consider experimental data on $\Upsilon_\alpha$. Unfortunately, nearly all the researches on glare have been conducted with $c_{\min}$ as criterion rather than $\alpha$. An exception is the experimental work of Cobb and Moss.[8] Their results have been compared with the curves of Fig. 4. The agreement is fair, but the data cover such a narrow range of $H_A/H_B$ that they do not constitute an adequate verification of Postulate I.

## 5. THRESHOLD, CONTRAST, AND VISUAL ANGLE FOR $H_A \leqq H_B$

It has long been recognized that glare is experienced when the helios of the surround is greater than the helios of the background. In recent years a similar effect has been discovered for dark surrounds. Cobb[18] in 1916 found that best vision is obtained when the helios of surround and background are equal. Lythgoe[9] (1932) discovered that visual acuity is reduced when the helios of the surround is too low. Luckiesh and Moss[19] (1939) used the blink test as criterion and concluded that best results are obtained when $H_S=H_B$.

The effect of dark surrounds is less noticeable than the effect of bright surrounds. In fact, the worst possible condition, which occurs with a completely dark surround, reduces visual acuity by only about 10 percent. Table III summarizes Lythgoe's data and shows that a change from optimum $H_S$ to a dark surround ($H_S\rightarrow0$) reduces visual acuity to 0.90 of its optimum value. This conclusion appears to be valid, yet the whole change is not greatly in excess of the experi-

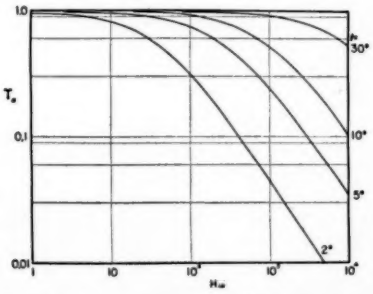



FIG. 7. Delos for a concentrated glare source in an otherwise uniform field. $H_B=H_s=100$ blondels. The glare source is at an angle $\theta$ from the line of sight and has a helios $H$ and solid angle $\omega$.

[18] P. W. Cobb, J. Exper. Psychol. **1**, 419 (1916).
[19] M. Luckiesh and F. K. Moss, Trans. I.E.S. **34**, 571 (1939).

mental error so that the shape of the curve cannot be accurately established.

Thus there appears to be no necessity for fitting a new curve to represent conditions with dark surrounds. The simplest assumption is that the curves of Fig. 4 are reflected unchanged about the vertical axis ($H_A/H_B=0$) to represent data obtained with dark surrounds. This condition is stated in—*Postulate II: When the adaptation helios $H_A$ is less than the background helios $H_B$, the Holladay principle and Postulate I apply but the ratio $H_A/H_B$ is replaced by $H_B/H_A$.*

Postulate II means that all the previous results, which were developed specifically for $H_A \geqq H_B$, can now be applied to the case of $H_A \leqq H_B$. Figures 3 and 4 are used just as with bright surrounds, but the former ratio $H_A/H_B$ is now replaced by $H_B/H_A$. The delos based on contrast becomes

$$\Upsilon_c=\left[\frac{K_1+1}{K_1+(H_B/H_A)^{\frac{1}{2}}}\right]^2, \quad H_A \leqq H_B, \quad (10a)$$

where $K_1=A_1H_B^{-\frac{1}{2}}$ as before. For most cases of room lighting, Eq. (10a) reduces to

$$\Upsilon_c=1/(H_B/H_A), \quad H_A \leqq H_B. \quad (11a)$$

Similarly, for visual angle,

$$\Upsilon_\alpha=\left[\frac{K_2+1}{K_2+(H_B/H_A)^{\frac{1}{2}}}\right]^2, \quad H_A \leqq H_B, \quad (16a)$$

where $K_2=A_2H_B^{-\frac{1}{2}}$. For high values of $H_B$, $\Upsilon_\alpha$ reduces to exactly the same equation obtained for $\Upsilon_c$. Therefore, under the usual conditions of interior-lighting practice, the delos is given unambiguously by

$$\Upsilon=1/(H_B/H_A), \quad H_A \leqq H_B. \quad (11b)$$

The extreme case of a dark surround is obtained when the surround has zero helios and comprises the whole visual field except for a central disk which is imaged on the fovea. If the diameter of this disk subtends a visual angle[2] of 0.0262 radian and there are no eye movements, the adaptation helios is

$$H_A=0.927H_B;$$

and from Eq. (11b), the delos is

$$\Upsilon=0.927.$$

Reference to Table III shows that Lythgoe obtained values of delos varying from 0.85 to 1.00 with a mean value of 0.90. Agreement is probably as satisfactory as can be obtained, because of the inherent variability of visual data.

## 6. APPLICATIONS

The general formulas and curves, applicable to any uniform or non-uniform helios distribution, are presented in the preceding sections.

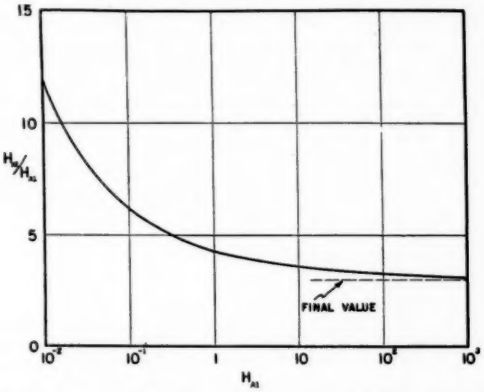



FIG. 8. Allowable helios ratio is a function of the adaptation helios $H_{A1}$ for the work. $\Upsilon=\frac{1}{3}$.

The results are given in terms of the adaptation helios, which must be calculated in each case by the methods that were developed in another paper.[2] Now consider a few examples to illustrate more definitely the application of the new methods.

In the first example, the observer is staring at a fixed background which is surrounded by a uniform field of helios $H_S$. The delos based on visual angle is calculated by means of Eq. (16), and the results are shown in Fig. 5. The curves for background helios of 10, 100, and $\infty$ are seen to be very nearly the same. Evidently if the helios of the surround is made sufficiently high, a considerable reduction in delos results. For $H_S=100H_B$, for instance, $\Upsilon=0.12$. In practice the effect would probably be even greater because of the difficulty of keeping the eyes fixed under such extreme conditions of helios variation. Figure 6 shows the same results for $H_B=100$ and includes dark surrounds.

A second example indicates the effect of a single glare source superimposed on a uniform

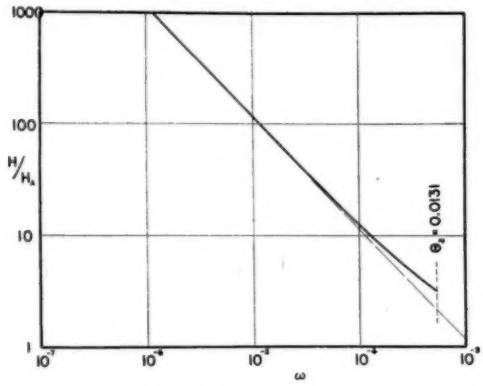



FIG. 9. Allowable helios $H$ of a glare source subtending a solid angle of $\omega$ steradians. $\Upsilon = \frac{1}{3}$.

field $H_S = H_B = 100$. The solid angle subtended by the source is $\omega$ and its helios is $H$. The angle from the line of sight to the glare source is $\theta$. Then the adaptation helios is

$$H_A = H_B + \frac{K \cos\theta}{\pi\theta^2}(H\omega). \tag{18}$$

Results plotted by means of Eq. (18) and Eq. (16) are given in Fig. 7.

These two illustrations show what can be done in specific cases. For other conditions, the adaptation helios is calculated by the more general formula. If there are $n$ concentrated glare sources, Eq. (3) is applicable. If the glare sources are larger in physical dimensions, integration is performed, as in the use of Eq. (4a).

### 7. ADAPTATION TO OTHER PARTS OF THE FIELD

The foregoing investigation is for a fixed line of sight. But ordinarily a person does not keep his eyes fixed on the work. A lighting system should be so designed that the observer can look up from his work without disturbing his adaptation too much. It seems likely that existing nonuniformity in lighting system is detrimental, not primarily because of its effect with fixed eye positions, but principally because it requires readaptation whenever the eyes are shifted. Thus the evaluation of a lighting system should depend upon transient adaptation. Transient conditions will be considered in Sections 8 and 9. Steady-state analysis is simpler, however, and gives results of practical importance.

Suppose that the observer looks at some object, such as a luminaire, until he becomes adapted to it. He then looks back at his work. At the initial instant, his adaptation remains at the previous level[1] and his ability to see the work can be evaluated by the methods of Section 6.

Let $H_{A1}$ be the steady-state adaptation helios when the eyes are directed at the work, and let $H_{A2}$ be the adaptation helios when the observer stares at any specified part of the visual field. According to Eqs. (13) and (14), the minimum visual angle when the eyes are directed at the work and adapted to it is

$$\alpha_1 = \frac{\alpha_\infty}{H_{B1}}[A_2 + H_{A1}^{\frac{1}{3}}]^3,$$

where $H_{B1}$ is the helios of the background of the work. When the eyes are directed at the work but are adapted to $H_{A2}$,

$$\alpha_2 = \frac{\alpha_\infty}{H_{B1}}[A_2 + H_{A2}^{\frac{1}{3}}]^3.$$

The relative visibility or *delos* for this case is equal to the ratio $\alpha_1/\alpha_2$, or

$$\Upsilon = \left[\frac{A_2 + H_{A1}^{\frac{1}{3}}}{A_2 + H_{A2}^{\frac{1}{3}}}\right]^3. \tag{19}$$

An allowable helios $H_{A2}$ can be determined from Eq. (19) if a limiting value of $\Upsilon$ be chosen. The equation may be written

$$H_{A2} \leqq \frac{1}{\Upsilon}[(1-\Upsilon^{\frac{1}{3}})A_2 + H_{A1}^{\frac{1}{3}}]^3. \tag{20}$$

Suppose it be specified that $\Upsilon$ shall not fall below $\frac{1}{3}$ on the instant when the observer looks back at his work. Then, since $A_2 = 0.412$,

$$H_{A2} \leqq 3[0.126 + H_{A1}^{\frac{1}{3}}]^3. \tag{20a}$$

TABLE IV. Allowable helios range for a spot in an otherwise uniform visual field. Calculated by means of Eq. (26).

| $\omega$ (steradian) | $H/H_{A1}$ Minimum | Maximum |
|---|---|---|
| $10^{-6}$ | 0.00086 | 1170 |
| $10^{-5}$ | 0.0085 | 118 |
| $10^{-4}$ | 0.079 | 12.7 |
| $5.4\times10^{-4}$ | 0.315 | 3.17 |
| Very large | 0.333 | 3.00 |

Equation (20a) may be utilized in computing the limiting value of adaptation helios for the brightest part of the visual field. It applies to any adaptation level $H_{A1}$. For the usual case where $H_{A1}^{\frac{1}{2}} \gg 0.126$, Eq. (20a) assumes the very simple form,

$$H_{A2}/H_{A1} \leqq 3. \qquad (20b)$$

Thus we arrive at an important criterion for room lighting: *No part of the visual field shall produce an adaptation helios greater than 3 times the adaptation helios for the work.*

Figure 8 shows the effect of varying the helios of the work, as computed by Eq. (20a). Ordinarily the area is large enough so that the adaptation helios $H_{A1}$ is essentially equal to the helios of the work. In reading, for instance, the eyes adapt themselves to the helios of the page. The curve of Fig. 8 indicates that at low values of $H_{A1}$, the helios of a large glare source can be as much as 12 times the adaptation helios. But as $H_{A1}$ increases, the allowable ratio decreases continuously and approaches 3 asymptotically. For values recommended for good lighting, $H_{A2}/H_{A1}$ is essentially equal to 3 as found in Eq. (20b).

Figure 8 exhibits a general relation that applies no matter how $H_{A1}$ and $H_{A2}$ are obtained. In reaching a physical picture of the results, however, we interpreted $H_{A1}$ as essentially equal to the helios of the work and $H_{A2}$ as equal to the helios of a luminaire or other bright part of the field. These approximations are practically true

TABLE V. Helios ratios to give sensation of shock. (Nutting, Trans. I.E.S. **11**, 939, 1916.)

| $H_A$ (blondel) | log $H/H_A$ Obs. J. B. | P. R. | P. G. N. | Mean |
|---|---|---|---|---|
| $10^{-5}$ | 7.81 | 7.45 | 6.78 | 7.35 |
| $10^{-1}$ | 4.50 | 4.64 | 4.65 | 4.60 |
| 1 | 4.18 | 3.74 | 3.78 | 3.90 |
| 10 | 3.30 | 3.30 | 3.30 | 3.30 |
| 100 | 2.62 | 2.79 | 2.76 | 2.72 |
| 2000 | 1.70 | 1.57 | 1.55 | 1.61 |
| 5760 | 1.30 | 1.33 | 1.35 | 1.33 |
| 81400 | 1.11 | 0.27 | 0.25 | 0.54 |

if both surfaces subtend large visual angles. Now consider another application. The glare source is in the form of a circular disk whose image does not fill the fovea. The remainder of the field is of a uniform helios equal to $H_{A1}$. Let the radius of the source subtend a visual angle of $\Theta_3$ radians ($\Theta_3 \leqq 0.0131$). Then, according to a previous paper,[2] the adaptation helios $H_{A2}$ produced by looking at the glare source of helios $H$ is

$$H_{A2} = \frac{0.923}{\Theta_1^2}[H\Theta_3^2 + H_{A1}(\Theta_1^2 - \Theta_3^2)]$$

$$+0.0192 H_{A1}\left[1 - \frac{\sin 2}{2} - Ci(2\Theta_1) + Ci(2)\right]. \qquad (21)$$

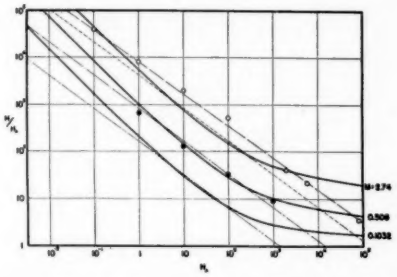


FIG. 10. Allowable helios ratio. ○ Experimental points of Nutting. ● Experimental points of Holladay, $\omega = 5.5 \times 10^{-2}$ steradian, sensation "just not unpleasant." —— Calculated curve of Nutting, Eq. (27).---- Calculated curve of Holladay, Eq. (28), $b = 2000$, boundary between "objectionable and intolerable." —·— Holladay, $b = 400$, "comfort-discomfort." —··— Holladay, $b = 80$, "limit of pleasure." ——— Theoretical curves, Eq. (38), based on photochemical theory. Very large glare source, cone vision.

If the constant $\Theta_1 = 0.0131$ is substituted,

$$H_{A2} = 5380[H\Theta_3^2 + H_{A1}(1.715 \times 10^{-4} - \Theta_3^2)]$$

$$+7.72 \times 10^{-2} H_{A1} \qquad (21a)$$

$$= 5380\Theta_3^2(H - H_{A1}) + H_{A1}.$$

Equation (21a) can be written

$$\frac{H}{H_{A1}} = \frac{1}{\Theta_3^2}\left[1.86 \times 10^{-4}\left(\frac{H_{A2}}{H_{A1}} - 1\right) + \Theta_3^2\right], \qquad (21b)$$

which applies to any ratio $H_{A2}/H_{A1}$. This ratio can be obtained from Fig. 8 for $\Upsilon = \frac{1}{3}$ or calculated for any other delos by means of Eq. (20). When $H_{A2}/H_{A1} \leqq 3$, Eq. (21b) becomes

$$H/H_{A1} \leqq (3.72 \times 10^{-4} + \Theta_3^2)/\Theta_3^2. \qquad (22)$$

For these small values of $\Theta_3$, the solid angle subtended by the glare sources is

$$\omega \leqq \pi\Theta_3^2.$$

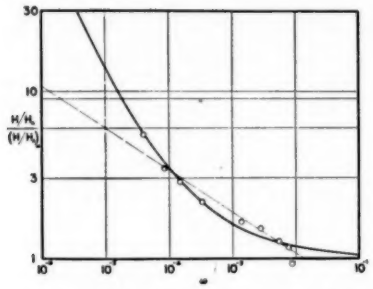



FIG. 11. Effect of size of glare source. ○ Experimental points of Holladay, $H_A=100$, sensation "perceptibly uncomfortable." — — Calculated curve of Holladay, Eq. (28). ——— Theoretical curve, Eq. (40), based on photochemical theory.

Thus Eq. (22) may be written

$$\frac{H}{H_{A1}} \leqq \frac{1.17\times10^{-3}}{\omega}+1 \qquad (23)$$

or

$$\frac{D}{H_{A1}} \leqq 3.72\times10^{-4}+\omega/\pi, \qquad (24)$$

where $D$ = luminous flux density (lumen $m^{-2}$) at the eye on a plane perpendicular to the line of sight.

Calculated values of $H/H_{A1}$ as a function of the solid angle $\omega$ are plotted in Fig. 9. At $\omega=10^{-6}$ steradian, the glare source may have 1170 times the helios of the work without producing a delos of less than $\frac{1}{3}$. But this ratio must be reduced rapidly as $\omega$ is increased. At $\omega=5.4\times10^{-4}$ steradian, which corresponds to a source whose image fills the fovea ($\Theta_3=\Theta_1=0.0131$ radian), the maximum allowable helios ratio is only 3.17. It will be noted from Fig. 9 that over most of the range of $\omega$, the curve is a straight line with slope equal to $-1$. In other words, for most purposes $(1.17\times10^{-3}/\omega)\gg1$ and Eq. (23) becomes approximately

$$H/H_{A1} \leqq 1.17\times10^{-3}/\omega. \qquad (23a)$$

Also,

$$D/H_{A1} \leqq 3.72\times10^{-4}. \qquad (24a)$$

The example considered in Eqs. (21) to (24) was made specific by limiting it to a single glare source in a uniform field of helios $H_{A1}$. The restriction that the source be a circular disk is not at all necessary, and the equations apply equally well to a source of any shape provided it is imaged entirely within the fovea. The requirement that the remainder of the field be uniform was introduced for simplicity. Equation (20) is perfectly general and can be employed with any known helios distribution. For instance, the case of a room lighted by a large number of diffusing-globe luminaires can be handled without undue difficulty by the methods of this paper.

The foregoing equations apply to $H_{A2} \geqq H_{A1}$. But the question may be how *dark* an object can be included in the field of view without too much reduction in visual acuity. The eyes are adapted to this darkest part of the field. They are then returned to the work, and the delos at the initial instant is evaluated. According to Postulate II, the previous equations will apply if the helios ratios are inverted. The general criterion is—*No part of the visual field shall produce an adaptation helios less than $\frac{1}{3}$ the adaptation helios for the work.*

Thus the criterion for any part of the visual field is

$$\tfrac{1}{3} \leqq H_{A2}/H_{A1} \leqq 3. \qquad (25)$$

For large areas, the values of adaptation helios are approximately equal to the values of helios for the surfaces. Thus the requirement may be stated that *no large surface in the room shall have a helios of more than 3 times or less than $\frac{1}{3}$ of the helios of the work.* For a small dark spot that does not fill the fovea, and an otherwise uniform field, Eq. (23) gives the criterion,

$$\frac{H}{H_{A1}} \geqq \frac{\omega}{1.17\times10^{-3}+\omega}.$$

The criterion for a small spot (bright or dark) is therefore

$$\frac{\omega}{1.17\times10^{-3}+\omega} \leqq \frac{H}{H_{A1}} \leqq \frac{1.17\times10^{-3}}{\omega}+1. \qquad (26)$$

Table IV gives some calculated values, showing how the limits tend to converge as the size of spot increases.

### 8. PREVIOUS RESEARCH ON SHOCK

The foregoing sections have dealt with steady-state conditions. Now turn to the consideration of transient conditions. The two principal researches are those of Nutting[10] and Holladay.[3]

In Nutting's tests, the observer became adapted to a helios $H_A$ by looking at a uniformly illuminated screen which filled most of his visual field. Suddenly the light source was switched off and a higher helios $H$ was substituted. The observer determined whether the transition to the new helios gave him an unpleasant shock or not.

The experimental results are given in Table V and Fig. 10. The allowable helios ratio ($H/H_A$) is obtained as a function of the adaptation helios $H_A$. The most striking feature is the great range of $H/H_A$. At $H_A=0.1$ blondel ($B=0.01$ ml), the glare source can have a helios almost 100,000 times that of the background without causing undue shock; but at the highest values of helios, the ratio must not be much greater than unity. Qualitative corroboration of this phenomenon is often given by reference to the sun and moon.[20] Since both bodies subtend essentially the same solid angle at the eye, the ratio between the helios of the full moon and the helios of the night sky is nearly the same as the ratio of the helios of the sun and the helios of the sky in the daytime. Yet in one case the effect is tolerable and in the other case it is glaring in the extreme. A less obvious example is found in interior lighting. Helios ratios that were entirely satisfactory in candle lighting or even lighting by carbon-filament lamps become very glaring when used at the higher levels employed in modern practice.[21]

The Nutting research is admirable in that it covers a very wide range of helios—from $10^{-5}$ to $10^5$ blondels, in fact. The data (Fig. 10) can be represented by a straight line, and thus Nutting advocated an empirical equation which in our notation is

$$H/H_A=8100H_A^{-0.68}. \qquad (27)$$

Obviously, this equation cannot continue to apply at extremely high values of $H_A$. The straight-line representation given by Eq. (27) would intersect the axis at a definite point, beyond which the eye would apparently be incapable of adaptation. Such conditions have never been found experimentally, even at higher helios than that for which the Nutting curve crosses the axis.[22] The idea also seems incompatible with the photochemical theory of adaptation,[23] which would lead one to expect the curve to approach the axis asymptotically as indicated by the heavy curves of Fig. 10.

Holladay[3] also investigated the subject, including as an additional factor the *size* of the glare source. The eyes were adapted to a large screen of uniform helios $H_A$. A circular opening in the center of the screen could be varied in size and helios. The spot, subtending a solid angle $\omega$ and having a helios $H$, was flashed on only momentarily. The observer appraised the shock of this sudden change.

Holladay's results are in approximate agreement with Nutting's, though the range (1 to 1000 blondels) is much less. Representative data obtained by Holladay are indicated by the black dots of Fig. 10. These data were obtained from his Fig. 26* and are for $\omega=5.5\times10^{-2}$ steradian and a sensation "just not unpleasant." The points seem to indicate a distinct upward concavity of the curve; though like Nutting, Holladay plotted his results as straight lines. The upper dotted curve of Fig. 10 represents the boundary between "objectionable and intolerable," the dot-dash curve represents the boundary between "comfort and discomfort," while the lowest curve exhibits the "limit of pleasure." The three curves are parallel and have a slope of $-0.30$, which is almost the same as that of the Nutting curve ($-0.32$). The absurdity of the linear representation is even more marked here than in the Nutting case. The comfort-discomfort curve cuts the axis at 14,000 blondels ($B=1400$ ml), which seems to say that even on a completely overcast day, conditions outdoors would be uncomfortably glaring.

The effect of size of glare source is shown in

TABLE VI. Constants in Holladay's equation.

| $b$ | Holladay constant $K$ | Designation |
|---|---|---|
| 10 | 0.3 | "Scarcely noticeable" |
| 80 | 1.2 | "At limit of pleasure" |
| 400 | 1.9 | "Boundary between comfort and discomfort" |
| 2000 | 2.6 | "Boundary between objectionable and intolerable" |

[20] Ward Harrison, Trans. I.E.S. **15**, 34 (1920).
[21] Ward Harrison, Trans. I.E.S. **32**, 208 (1937).
[22] H. Eguchi, C.I.E. Proc., 1932, p. 36; J. Steinhardt, J. Gen. Physiol. **20**, 185 (1936); Hecht, Haig, and Chase, J. Gen. Physiol. **20**, 831 (1937).
[23] S. Hecht, J. Gen. Physiol. **18**, 767 (1935).
* Reference 5, p. 306.

Fig. 11. The points were taken from Holladay's Fig. 25. He represents them by the straight line, which obviously does not give a very good fit. The complete Holladay equation is

$$H = bH_A^{0.30}(1/\omega)^{0.25}$$

or

$$H/H_A = bH_A^{-0.70}(1/\omega)^{0.25}, \tag{28}$$

where $\omega$ is the solid angle (steradian) subtended by the spot, $H$ is the helios (blondel) of the spot, $H_A$ is the adaptation helios, and $b$ is a constant (Table VI).

The principal objections to the Holladay results on the size of glare source are the small range of $\omega$ ($10^{-2}$ to $10^{-4}$ steradian)—which makes any extrapolation to such values as $\omega = 1$ very doubtful—and the poor fit of the empirical equations as shown in Fig. 11. Theoretical considerations also show that Eq. (28) cannot be correct with respect to $\omega$. If the angle is so small that the concentrated source cannot be resolved by the eye, its visual effect depends on the product $H\omega$. This is Riccò's law which is found to hold not only for very small sources but for sources approaching the visual angle subtended by the fovea (0.0131 radian). Thus at small angles, the term $(1/\omega)$ in Eq. (28) must appear in the first power, not in the 0.25 power. Moreover, at sufficiently large visual angles the effect of size becomes practically negligible. Thus the Nutting-Holladay equation is unsatisfactory and its use may lead to gross errors.

### 9. APPLICATION OF THE PHOTOCHEMICAL THEORY

The next step is to examine the experimental results of Nutting and Holladay in the light of the photochemical theory of vision.[23] According to this theory, a sensitive substance $S$ in the cones is broken down into other substances by the action of light, or

$$S + h\nu = A + B.$$

Let $x$ = number of molecules of $A$ in one cone, expressed as a fraction of the maximum possible number. Then $(1-x)$ will represent the relative concentration of $S$. The velocity of the photochemical reaction depends on $H$ and $x$. For a fixed pupil size,

$$\left(\frac{dx}{dt}\right)_1 = H \cdot \Phi_1(x). \tag{29}$$

There is also a chemical reaction by which $A$ and $B$ combine to form $S$. The reaction velocity will be denoted by $(dx/dt)_2$. The complete reaction velocity is then

$$\frac{dx}{dt} = \left(\frac{dx}{dt}\right)_1 + \left(\frac{dx}{dt}\right)_2. \tag{30}$$

Equation (30) represents a differential equation whose solution gives the concentration $x$ as a function of time, helios, etc. For equilibrium, the solution is[24]

$$H_A = Kx_A/\Phi_1(x_A), \tag{31}$$

where $H_A$ = adaptation helios (blondel), $x_A$ = equilibrium concentration corresponding to $H_A$, $K$ = constant.

The studies of Hartline, Graham,[25] and others have shown that the impulses in a single fiber of the optic nerve are constant in amplitude but vary over a wide range of frequency, depending on the retinal illumination. Thus it seems reasonable to assume that the sensation depends on the number of nerve impulses per second, other factors being equal. The frequency of nerve impulses is assumed[24] to be

$$f = k_3\left(\frac{dx}{dt}\right)_1 = k_3 H \Phi_1(x). \tag{32}$$

We now apply the foregoing theory to the experiments of Nutting and Holladay. Suppose that the eye is adapted to a helios $H_A$ and that the helios is suddenly raised to $H$. The eye will immediately start adapting itself to the new helios, but at the beginning there will be a burst of nerve impulses[25] caused by the sudden increase in helios. If the frequency of impulses in this burst is sufficiently high, an unpleasant sensation of glare results. We assume that the shock

[24] P. Moon and D. E. Spencer, J. Opt. Soc. Am. **35**, 43 (1945); *Visual dark adaptation: a mathematical formulation*, to be published.

[25] H. K. Hartline and C. H. Graham, J. Cell. and Comp. Physiol. **1**, 277 (1932); C. H. Graham, *Handbook of General Experimental Psychology* (Worcester, 1934), p. 829.

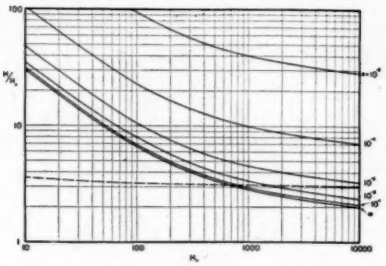


FIG. 12. Recommended helios ratio. $H_A$ is the adaptation helios (blondel) and $H$ is the helios of a bright spot subtending the solid angle $\omega$ (steradian). — — Steady-state, very large spot, Eq. (20a). ——— Transient state, Eqs. (38) and (40), $M=0.1032$.

depends on the *difference in frequency* $(f-f_A)$, where $f_A$ = equilibrium frequency corresponding to $H_A$ and $f$ = burst frequency caused by helios $H$. According to Eq. (32),

$$f-f_A=k_3(H-H_A)\Phi_1(x_A). \tag{33}$$

Consider a specific adaptation helios $H_{AN}$, and suddenly change to a value $H_N$ which gives a barely perceptible shock. A norm is thus established to which all other conditions can be compared. The normal frequency is

$$f_N-f_{AN}=k_3(H_N-H_{AN})\Phi_1(x_{AN}). \tag{34}$$

To find the helios $H$ that will produce the same degree of shock when suddenly applied to the eye adapted to $H_A$, we merely equate frequencies, or

$$f-f_A=f_N-f_{AN},$$

$$(H-H_A)\Phi_1(x_A)=(H_N-H_{AN})\Phi_1(x_{AN})$$

or

$$(H/H_A-1)=\frac{H_N/H_{AN}-1}{H_A/H_{AN}}\frac{\Phi_1(x_{AN})}{\Phi_1(x_A)}. \tag{35}$$

But for equilibrium, Eq. (31) gives

$$H_A=Kx_A/\Phi_1(x_A), \quad H_{AN}=Kx_{AN}/\Phi_1(x_{AN}).$$

Thus,

$$\frac{1}{H_A/H_{AN}}\frac{\Phi_1(x_{AN})}{\Phi_1(x_A)}=\frac{x_{AN}}{x_A}. \tag{36}$$

Substitution in Eq. (35) results in the equation,

$$(H/H_A-1)=(x_{AN}/x_A)(H_N/H_{AN}-1). \tag{37}$$

For simplicity, the foregoing derivation has been for a fixed pupil. But if the natural pupil is employed, it will have a different diameter for $H_A$ and for $H_{AN}$. Thus the Stiles factor must be introduced,[26] and Eqs. (33) and (34) become

$$f-f_A=k_3\zeta(\delta_A, 2)[H-H_A]\Phi_1(x_A),$$

$$f_N-f_{AN}=k_3\zeta(\delta_{AN}, 2)[H_N-H_{AN}]\Phi_1(x_{AN}).$$

Equation (37) then assumes its final form,

$$\frac{H}{H_A}=1+\frac{M}{\zeta_A x_A}, \tag{38}$$

where the constant $M$ is

$$M=\zeta_{AN}x_{AN}(H_N/H_{AN}-1). \tag{39}$$

To obtain a criterion for best lighting conditions, take $H_N/H_{AN}=3$ for $H_A=800$ blondels. This assumption gives a ratio $H/H_A$ that is comparable with that obtained in the steady-state analysis of the preceding section. From previous work, the concentration for the cones[24] is $x_{AN}=0.030$ and the Stiles factor[26] is $\zeta_{AN}=1.72$. Thus

$$\mathbf{M=0.1032.}$$

The curve plotted by use of this constant in Eq. (38) is shown in Fig. 10. It lies between the Holladay curves "limit of pleasure" and "comfort-discomfort." We recommend this curve for use where the best lighting is desired.

The curve $M=0.508$ of Fig. 10 was calculated by means of Eq. (38) to see how close the Holladay data could be approximated. The new curve is seen to agree at least as closely with the four experimental points as the straight line did. A third curve is plotted for $M=2.74$. Its approximation to the Nutting data is not very good. But, as was pointed out in Section 8, the Nutting data cannot be correct at either end of the scale. At low helios, the slope of the curve must be $-1.0$, not $-0.32$ as Nutting found; and at high helios, the slope must approach zero. Both of these requirements are met by the new curves based on photochemical theory.

The effect of *size of source* is also handled analytically. If $(H/H_A)_\infty$ is the allowable helios ratio with very large illuminated surfaces (Fig.

[26] P. Moon and D. E. Spencer, J. Opt. Soc. Am. **34**, 319 (1944).

10), then the allowable helios ratio for smaller sources is given by the expression,[24]

$$\frac{H}{H_A}=\frac{(H/H_A)_\infty}{\omega}[A_3+\omega^{\frac{1}{2}}]^2, \qquad (40)$$

where $\omega$ = solid angle (steradian) subtended by the spot having helios $H$,

$$A_3 = 0.00874 \text{ for cone vision.}$$

The heavy curve of Fig. 11 is plotted by means of Eq. (40) and is a much better fit to the Holladay data than the straight line.

A practical application of the photochemical theory is the determination of allowable helios ratios in room lighting. Ordinarily the observer will be adapted to the work so that $H_A$ is essentially equal to the helios of the work. He then glances at some other part of the room and experiences an unpleasant shock if the helios ratio is too great. The allowable helios ratio is calculated from Eqs. (38) and (40) and is plotted in Fig. 12.

As an example of the use of Fig. 12, suppose that the observer is reading a book with $H_A = 300$ blondels, how high a helios is permissible in a luminaire subtending a solid angle of $10^{-2}$ steradian (1 sq. ft. at 10 ft.)? Reference to Fig. 12 shows that the allowable ratio is 4.8, or the allowable helios of the luminaire is 1840 blondels.

The photochemical theory can also be applied to the criterion of the darkest allowable object in the field of view. The worst case will obtain when the observer looks away from his work for sufficient time to adapt to the darkest part of the field. This adaptation helios is calculated by the methods of Section 3. If the observer now looks back at the work, he may experience a shock. The allowable helios ratio is again obtained from Eq. (38) or Fig. 12.

## 10. SUMMARY

The paper presents an analysis of the effect of non-uniform visual fields. Three conditions are considered: (a) Eyes fixed on the work, effect of non-uniform surround is evaluated. (b) Steady-state adaptation to any part of the visual field, effect of suddenly looking back at work is evaluated. (c) Transient conditions, criterion is the unpleasant sensation rather than the reduction in vision.

Condition (a) is a necessary step in the study of (b). It is of practical importance in some cases such as glare from automobile headlamps but does not seem to be of much importance in the establishment of a criterion for room lighting.

Condition (b) leads to the important theoretical conclusion that for best vision *no part of the field should produce an adaptation helios more than 3 times or less than $\frac{1}{3}$ the adaptation helios of the work.*

For large surfaces, the values of adaptation helios are practically equal to the values of helios for the surfaces. For small surfaces, the actual helios ratio can be greater than 3 without violating the above criterion. Equations and graphs allow calculation under all conditions.

Condition (c) makes use of a recently developed photochemical theory of vision. The theoretical curves are in fairly good agreement with the experimental work of Nutting and of Holladay. Both theory and experiment indicate that very large helios ratios may be employed safely at low helios levels. Over the range of good interior lighting, however, the curves are fairly flat (Fig. 12) and tend to corroborate the criterion of (b).